# ALTE 6th INTERNATIONAL CONFERENCE

BOLOGNA, ITALY | 2017

# Learning and Assessment: Making the Connections

3–5 May 2017

## CONFERENCE PROCEEDINGS

ISSN 2789-2344

ALTE
Association of Language Testers in Europe

In collaboration with

cliQ
CERTIFICAZIONE LINGUA ITALIANA DI QUALITA'

and supported by

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

ALTE

Association of Language Testers in Europe

Association of Language Testers in Europe, 1 Hills Road, Cambridge CB1 2EU, United Kingdom

www.alte.org

If you would like to quote any part of this work, please use the following details:

ALTE (2017). Learning and Assessment: Making the Connections – Proceedings of the ALTE 6th International Conference, 3-5 May 2017.

The opinions expressed in this work are those of the authors and do not necessarily reflect the views of ALTE. ALTE has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

# Learning and Assessment:
# Making the Connections

ALTE 6th International Conference, 3-5 May 2017

# CONFERENCE PROCEEDINGS

**Editor/Production Lead**

Esther Gutiérrez Eugenio, ALTE Secretariat Manager

**Production Team**

John Savage, Publications Assistant, Cambridge English Language Assessment

Mariangela Marulli, ALTE Secretariat Coordinator

Alison French, Consultant

**Acknowledgements**

We would like to extend our most sincere thanks to all the contributors who have diligently replied to all our queries (even during the summer months) and thus allowed this publication to come to life on such a tight schedule.

# Conference Proceedings

ALTE 6th
INTERNATIONAL CONFERENCE
BOLOGNA, ITALY | 2017

## Learning and Assessment: Making the Connections

… in a globalised economy ........................................ 1

La Verifica come Occasione di Apprendimento e Aggiornamento Attraverso l'Esperienza della Certificazione Glottodidattica DILS-PG di II Livello ........................................ 2

Certification of Proficiency in Polish as a Foreign Language and its Influence over the Polish Labour Market ........................................ 12

Beliefs Driving the Assessment of Speaking: An Empirical Study in a Brazilian Public Classroom ........................................ 20

Assessment in a Globalized Economy: A Task-based Approach to Assess the Proficiency of Dutch in Specific Occupational Domains ........................................ 26

Implications of Employing Performance-based Testing in a University Context ........................................ 32

Testing Pre-service Teachers' Spoken English Proficiency ........................................ 39

Validating University Entrance Policy Assumptions. Some Inconvenient Facts. ........................................ 46

Testing the Test: How Politics Influenced the Reception of an English Test for Lecturers ........................................ 51

Empowering Learners for a Demanding Labor Market: The "Groups for Experimentation of Plurilingualism" Education Program in Catalonia ........................................ 57

Hong Kong Diploma of Secondary Education (HKDSE) English Language Level Descriptors: Stakeholder Recognition and Understanding ........................................ 64

Academic Literacy and Language Proficiency in Testing: Overlapping and Diverging Constructs ........................................ 73

… in a multilingual world ........................................ 86

Mediation and Exploiting One's Plurilingual Repertoire: Exploring Classroom Potential with Proposed New CEFR Descriptors ........................................ 87

QualiCEFR: A Quality Assurance Template to Achieve Innovation and Reform in Language Education through CEFR Implementation ........................................ 96

Tatar Language Tests for Maintaining a Language of a Minority Ethnic Group in Russia: The Case of Kazan ........................................ 104

The Use of Test Taker Productions in Redesigning Writing Assessment Grids. A Corpus Based Study ........................................ 112

The Evaluation of Chinese Students of Italian L2 in China and in Italy: Practices at the Universities for Foreigners of Perugia and Siena ........................................ 118

Verifica delle Possibilità per una Certificazione della Lingua Araba ........................................ 124

The Implementation of a French Language Certification: Positive Washback and Wider Resulting Effects ........................................ 129

Instructors' Perceptions of the Construct-Relevance of Language in the Assessment of Literature ........................................ 136

Diagnostic Assessment: Incorporation of Assessment and Teaching in Foreign Language Education ........................................ 141

Reconsidering The Impact of Language Assessment on Language Learning and Teaching: A Survey on an Italian Examination for Young Learners ........ 147
Investigating Scoring Procedures in Language Testing ........ 154
A Comparative Study on the Washback of CET-6, IELTS and TOEFL iBT Writing Tests: Evidence from Chinese Test-takers' Perspectives ........ 160
The Generation of an Individualized Cognitive Diagnostic Report for College English Writing ........ 166
Assessment in the future: A Cognitive Diagnostic Modelling for College English Reading Test ........ 172
The Assessment Literacy of Language Teachers: A Case Study of Teachers of Portuguese as A Foreign Language 179
Comparing Speaking Performances Across Tests and Languages: Evaluating the Success of an institutional Rater-training Program ........ 194
Exploring Teachers' Language Assessment Literacy: A Social Constructivist Approach to Understanding Effective Practices ........ 201
Construction of Attribute-based Dynamic Mediation for Cognitive Diagnostic EFL Listening Test ........ 208
Integrating Corpus Linguistics and Classroom-based Assessment: Evidence from Young Learners' Written Corpora. 214
Cut Scores for Combined Constructs ........ 219
Towards a Scale of Academic Language Proficiency ........ 224
Developing a Japanese Language Test for a Multilingual Online Assessment System: Towards an Action-oriented Approach to Japanese Instruction in Europe ........ 236
How to Assess Mediation? ........ 245

… in the digital era ........ 252

The Impact of the Integrated Teaching, Learning and Assessment Framework on Students' Writing Perceptions and Performance ........ 253
Análisis de la Severidad de los Calificadores de una Prueba de Expresión Escrita en el Contexto de una Prueba de Dominio ........ 261
Predicting Readability of Texts for Italian L2 Students: A Preliminary Study ........ 272
The Impact of On-line Teaching Practices on Greek EFL Learners' Reading Perceptions & Exam Performance ........ 279
Integrating Technology with Language Assessment: Automated Speaking Assessment ........ 286
Using Technologies for Teaching a Subject with CLIL: A Look at Teachers' Reactions ........ 292

# Language learning,
# teaching and assessment…

## … in a globalised economy

ALTE
Association of Language Testers in Europe

# La Verifica come Occasione di Apprendimento e Aggiornamento Attraverso l'Esperienza della Certificazione Glottodidattica DILS-PG di II Livello

**Nicoletta Santeusanio**, Università per Stranieri di Perugia – CVCL, Italia

**Abstract:** Obiettivo del presente contributo è quello di evidenziare, attraverso l'esperienza della certificazione in "Didattica dell'Italiano Lingua Straniera" DILS-PG di II livello dell'Università per Stranieri di Perugia, come la verifica delle conoscenze e competenze richieste a un docente di italiano a stranieri per poter svolgere al meglio il proprio lavoro non sia finalizzata solo alla valutazione, ma possa anche costituire un'occasione di apprendimento, sistematizzazione, riflessione sulla propria esperienza di insegnamento e uno stimolo per un aggiornamento continuo.

Verranno, pertanto, presentati i risultati preliminari di una ricerca più vasta, basata su una sistematica raccolta e analisi dei dati che emergono dalla somministrazione della DILS-PG di II livello. Dopo aver descritto il profilo degli iscritti all'esame, relativamente al campione di riferimento, verranno analizzate le risposte dei candidati nel questionario somministrato al termine della prova, soffermandosi sui singoli casi e proponendo elementi e spunti di riflessione in base alle risposte stesse.

## 1 Introduzione

Nel presente contributo verrà presentata l'esperienza della certificazione in "Didattica dell'Italiano Lingua Straniera" DILS-PG di II livello, elaborata e prodotta dal CVCL (Centro per la Valutazione e le Certificazioni Linguistiche) dell'Università per Stranieri di Perugia, con l'obiettivo di dimostrare come la verifica delle conoscenze e competenze richieste a un docente di italiano lingua non materna per poter insegnare in maniera efficace non sia finalizzata solo alla valutazione, ma possa anche rivelarsi un'occasione di apprendimento, sistematizzazione, riflessione sulla propria esperienza di insegnamento e uno stimolo per un aggiornamento continuo.

Verrà pertanto descritto il profilo dei candidati all'esame DILS-PG di II livello, relativamente al campione di riferimento, e verranno analizzate le risposte degli stessi alle domande presenti nel questionario somministrato al termine della prova. Si tratta di uno studio che è parte integrante di una ricerca più ampia, basata sulla sistematica raccolta e analisi dei dati che provengono dalla somministrazione degli esami DILS-PG (Marasco/Santeusanio, 2016, Santeusanio *in corso di stampa* a e b). In questa sede ne illustreremo i risultati preliminari con un'attenzione particolare ai singoli casi e la proposta di spunti di riflessione in base alle risposte fornite dai candidati.

## 2 La certificazione in "Didattica dell'Italiano Lingua Straniera" (DILS-PG)

La certificazione in "Didattica dell'Italiano Lingua Straniera" (DILS-PG) è una certificazione glottodidattica articolata in due livelli: 1) DILS-PG di base di I livello 2) DILS-PG di II livello. È rivolta a docenti italiani e stranieri di italiano lingua non materna ed è specifica per l'accertamento delle conoscenze e competenze glottodidattiche[1] necessarie per intraprendere la professione di docente di italiano a stranieri o per il riconoscimento del servizio prestato. I due livelli certificatori rimandano a profili di docenti con conoscenze e competenze diverse: il primo è

[1] In generale sulla formazione e sulle competenze dei docenti di italiano L2 si rimanda a Ciliberti, 2007, Jafrancesco, 2007, a Diadori, 2010. Sulla formazione all'Università per Stranieri di Perugia cf. Santeusanio, 2013b.

rivolto a insegnanti che hanno un'esperienza limitata in termini di ore di insegnamento e circoscritta ai livelli A1-B1 del *Quadro comune europeo di riferimento* (*QCER*), il secondo a insegnanti che hanno un'esperienza maggiore ed estesa a tutti i livelli del *QCER* (https://www.unistrapg.it/it/node/1782). I saperi testati nei due livelli, pur essendo verificati con un grado di approfondimento diverso e con metodi diversi (Marasco & Santeusanio, 2016, p.114), sono il sapere, il saper riflettere e il saper fare. Essi vengono accertati all'interno di tre fascicoli: il primo dedicato alle conoscenze teoriche di glottodidattica e relative alla metalingua, il secondo alla consapevolezza metodologico-didattica con l'analisi dei materiali didattici e l'osservazione della classe e il terzo alle capacità operative relative non solo alla progettazione e costruzione di attività didattiche, ma anche alla gestione della classe e all'uso delle tecnologie e di Internet (Santeusanio, 2014b, pp.7-14). Si tratta di saperi trasversali all'insegnamento delle lingue straniere che compaiono anche all'interno dell'*European Profiling Grid* (*EPG*) suddivisi in base a tre fasi di sviluppo della figura del docente di lingue (http://egrid.epg-project.eu/it/egrid; Rossner, 2010).

## 3 Descrizione del profilo dei candidati

Prima di descrivere il profilo dei candidati all'esame DILS-PG di II livello, ricordiamo che possono iscriversi sia i laureati che i diplomati (Santeusanio, 2014a, p.5). Nel caso dei laureati, possono accedere all'esame non solo coloro che hanno una laurea specifica per l'insegnamento dell'italiano a stranieri ma anche coloro che hanno una laurea generica purché abbiano un'esperienza minima di insegnamento di italiano a stranieri o di altra lingua straniera di 400 ore certificate o, in alternativa, un titolo *post lauream* specifico per l'insegnamento dell'italiano lingua non materna (master o scuola di specializzazione). Anche ai diplomati è consentito l'accesso all'esame a condizione che abbiano un'esperienza di almeno 1500 ore certificate (Santeusanio, 2013a, p. 68).

I dati che analizzeremo sono relativi ai candidati che hanno sostenuto l'esame nella sessione di settembre 2016, dei quali ci sono pervenute 101 schede informative, e a quelli della sessione di febbraio 2017[2], dei quali abbiamo ricevuto 112 schede.

La maggior parte dei candidati ha un'età compresa tra i 31 e 50 anni (63%), è laureata (più del 90%) e tra le lauree indicate prevale quella in Lingue e letterature straniere (43% nel 2016 e 38% nel 2017). Per quanto riguarda la formazione, il 53% ca. nel 2016 e il 63% nel 2017 ha una formazione in didattica dell'italiano a stranieri o di altre lingue straniere, mentre per quanto concerne l'aggiornamento, solo il 31% nel 2016 e il 38% nel 2017 ha frequentato corsi in didattica dell'italiano a stranieri o di altre lingue straniere. Inoltre il 43% nel 2016 e il 40% nel 2017 ha più di 5 anni di esperienza[3], alcuni addirittura hanno più di 15 anni di insegnamento (11% nel 2016 e 16% nel 2017). Il contesto di insegnamento è molto variegato, tra le realtà indicate compaiono le seguenti: scuole pubbliche o paritarie, cooperative o associazioni che si

[2] D'ora in poi per ragioni di brevità si farà riferimento al 2016 per la sessione di settembre 2016 e al 2017 per la sessione di febbraio 2017. Quando sarà necessario, verrà opportunamente indicato il mese per distinguerlo da altre sessioni che hanno avuto luogo nel 2016.

[3] La percentuale potrebbe essere più alta dal momento che molti candidati si sono limitati a barrare la casella "più di 1 anno" senza specificare il numero effettivo di anni.

occupano dell’accoglienza dei migranti, ex-CTP ora CPIA, scuole private di lingue sia in Italia che all’estero, Università, SLEE (Scuola di Lingue Estere dell’Esercito), sindacati, agenzie formative, ecc. Significativo è il dato relativo a coloro che insegnano nelle scuole pubbliche o paritarie di ogni ordine e grado (44% nel 2016 e 40% nel 2017) e il fatto che tra di esse prevalgano quelle superiori di II grado (44% nel 2016 e 37% nel 2017). L’aumento della percentuale di coloro che insegnano nelle scuole pubbliche e paritarie si era registrato già nella sessione di febbraio 2016, passando dall’8% del febbraio 2014 al 43% del febbraio 2016 (Santeusanio *in corso di stampa* b), in seguito all’annuncio nell’estate del 2015 dell’istituzione della nuova classe di concorso A23 per l’insegnamento dell’italiano agli alunni stranieri e al riconoscimento, da parte del Ministero dell’Istruzione, dell’Università e della Ricerca (MIUR) con il D.M. 92 del 23 febbraio 2016, della DILS-PG di II livello tra i titoli specifici per l’affidamento di incarichi connessi all’insegnamento dell’italiano lingua non materna e all’inserimento della stessa nel D.M. 94 tra i titoli valutabili nel concorso a cattedre per docenti della scuola pubblica. Per quanto riguarda le lingue insegnate, la maggior parte dei candidati dichiara di insegnare italiano a stranieri (o in classi con un’alta percentuale di alunni stranieri): per alcuni è l’unica lingua straniera insegnata (37% nel 2016 e 43% nel 2017), per altri è una delle lingue straniere insegnate (30% nel 2016 e 25% nel 2017).

All’interno della scheda informativa i candidati dovevano specificareanche i motivi che li avevano spinti a iscriversi all’esame: il motivo più indicato sia nel 2016 (22%) che nel 2017 (24%) è stato ‘crescere come insegnante’. È da notare che nel 2016 anche un altro aspetto, ovvero la mancanza di un titolo specifico, ha avuto la stessa percentuale, probabilmente in seguito all’emanazione dei due decreti su menzionati e all’istituzione della nuova classe di concorso.

## 4 Analisi dei questionari

Al termine della prova viene somministrato un questionario per raccogliere il feedback dei candidati sull’esame. Nelle due sessioni che in questo contributo intendiamo analizzare, al questionario standard sono state aggiunte due domande più specifiche relative appunto alla DILS-PG come occasione di apprendimento, sistematizzazione, riflessione sulla propria esperienza di insegnamento e stimolo per un aggiornamento continuo. I candidati dovevano esprimere una valutazione da 1 a 4 in relazione agli aspetti elencati. Si riportano qui di seguito le due domande presenti nel questionario:

**1) Quanto la preparazione all’esame DILS-PG di II livello Le è stata utile in relazione ai seguenti aspetti?  Dia una valutazione da 1 a 4.**

Poco ---------------------------------→ molto

1 2 3 4

| | | | | |
|---|---|---|---|---|
| Sistematizzare le Sue conoscenze relative all’insegnamento di una L2 | 1 | 2 | 3 | 4 |
| Approfondire le Sue conoscenze relative all’insegnamento di una L2 | 1 | 2 | 3 | 4 |
| Riflettere sulla Sua esperienza come docente di una L2 | 1 | 2 | 3 | 4 |
| Acquisire consapevolezza del perché di certe scelte didattiche | 1 | 2 | 3 | 4 |
| Rivedere il Suo modo di insegnare una L2 | 1 | 2 | 3 | 4 |
| Avere nuovi stimoli per la Sua attività di docente di una L2 | 1 | 2 | 3 | 4 |

**2)  Dopo essersi preparato/a all’esame DILS-PG di II livello quanto pensa di riconsiderare il Suo modo di insegnare in relazione ai seguenti aspetti?  Dia una valutazione da 1 a 4.**

Poco ---------------------------------→ molto

1 2 3 4

| | | | | |
|---|---|---|---|---|
| Scegliere in maniera critica il manuale da adottare in classe | 1 | 2 | 3 | 4 |
| Creare autonomamente materiale didattico per i Suoi studenti | 1 | 2 | 3 | 4 |
| Prestare attenzione alla lingua utilizzata in classe | 1 | 2 | 3 | 4 |
| Prestare attenzione alla formulazione delle istruzioni | 1 | 2 | 3 | 4 |
| Migliorare nella gestione di eventuali problemi che si possono verificare in classe favorendo anche la partecipazione attiva degli studenti | 1 | 2 | 3 | 4 |
| Scegliere con consapevolezza le modalità di correzione degli errori commessi dagli studenti in relazione alle circostanze e/o alle finalità dell’attività svolta | 1 | 2 | 3 | 4 |
| Riflettere sul ruolo delle tecnologie e di Internet nell’insegnamento di una L2 | 1 | 2 | 3 | 4 |

Altro …………………………………………………………………………………………………………

Per quanto riguarda la sessione di settembre 2016 sono pervenuti 91 questionari, dei quali solo 58 (64%) risultavano compilati anche in relazione alle domande su menzionate, in 27 casi (30%) i candidati non avevano risposto né alla domanda 1 né alla domanda 2, in 5 casi (5%) non avevano risposto alla domanda 1 e in un caso (1%) alla domanda 2. Per quanto concerne, invece, la sessione di febbraio 2017, sono stati rispediti dalle sedi d'esame 111 questionari, dei quali 87 (78%) risultavano compilati relativamente alle due domande in questione. Inoltre, in 15 casi (14%) i candidati non avevano fornito risposte ad entrambe le domande, in 7 casi (6%) non avevano risposto alla domanda 1 e in 2 casi (2%) alla domanda 2.

Se analizziamo ora le valutazioni espresse dai candidati in relazione alle due domande e agli aspetti riportati nel questionario, notiamo, per quanto riguarda la domanda 1, che sia nel 2016 sia nel 2017 i candidati hanno assegnato in prevalenza il valore '4' a quasi tutti gli aspetti, che nel 2017 hanno espresso in media valori più alti e che l'unico aspetto a cui è stato attribuito il valore '3' in percentuale più alta sia nel 2016 che nel 2017 è stato 'rivedere il proprio modo di insegnare' con un conseguente aumento della percentuale dei valori '1' e '2' (10% e 27% nel 2016 e 11% e 19% nel 2017). Inoltre, risulta che l'aspetto che ha riportato la percentuale più alta per il valore '4', sia nel 2016 che nel 2017, è stato 'avere nuovi stimoli per la propria attività di docente di una L2'.

Si considerino a tal proposito i grafici 1 e 2 e la tabella 1 di seguito in cui sono state riportate le percentuali relative ai valori '3' e '4'.

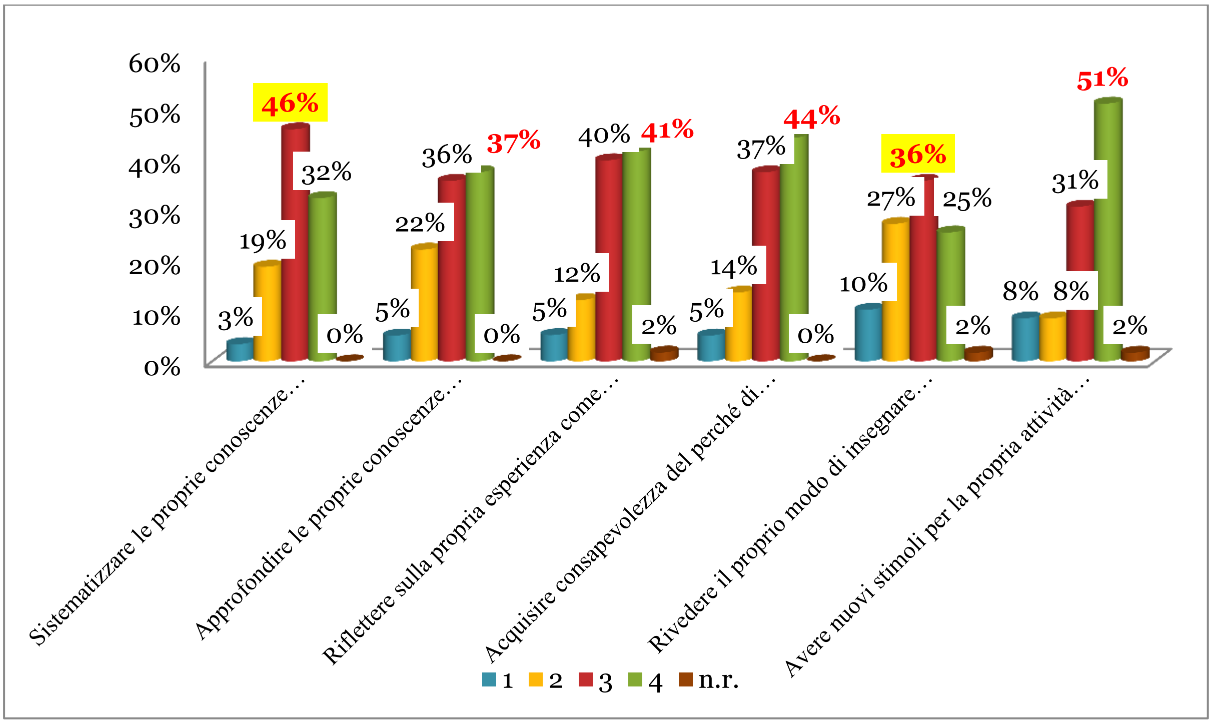


**Grafico 1.** Valori attribuiti dai candidati della sessione di settembre 2016 agli aspetti indicati nella domanda 1 (Quanto la preparazione all'esame DILS-PG di II livello Le è stata utile in relazione ai seguenti aspetti? Dia una valutazione da 1 a 4).

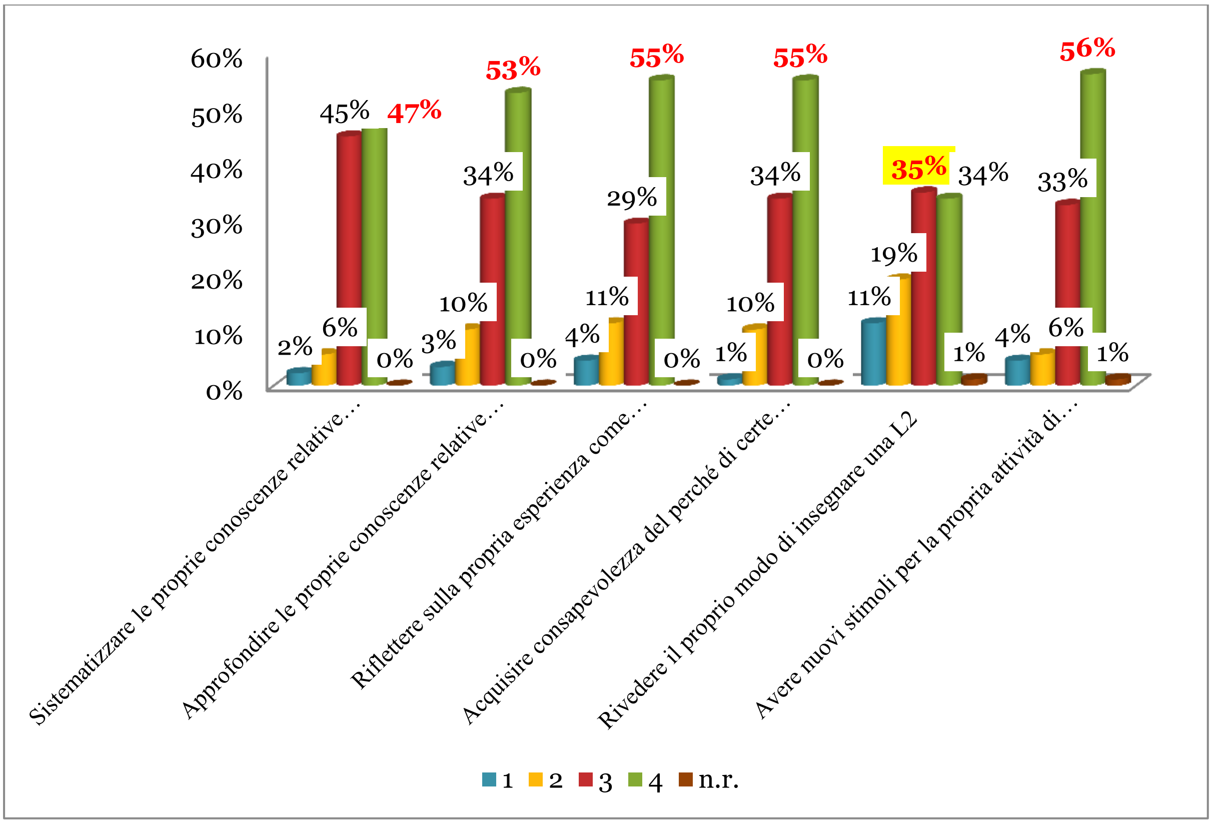


**Grafico 2.** Valori attribuiti dai candidati della sessione di febbraio 2017 agli aspetti indicati nella domanda 1 (Quanto la preparazione all'esame DILS-PG di II livello Le è stata utile in relazione ai seguenti aspetti? Dia una valutazione da 1 a 4).

| | 2016 | | 2017 | |
|---|---|---|---|---|
| | 3 | 4 | 3 | 4 |
| Sistematizzare le proprie conoscenze relative all'insegnamento di una L2 | **46%** | 32% | 45% | **47%** |
| Approfondire le proprie conoscenze relative all'insegnamento di una L2 | 36% | **37%** | 34% | **53%** |
| Riflettere sulla propria esperienza come docente di una L2 | 40% | **41%** | 29% | **55%** |
| Acquisire consapevolezza del perché di certe scelte didattiche | 37% | **44%** | 34% | **55%** |
| Rivedere il proprio modo di insegnare una L2 | **36%** | 25% | **35%** | 34% |
| Avere nuovi stimoli per la propria attività di docente di una L2 | 31% | **51%** | 33% | **56%** |

**Tabella 1.** I valori '3' e '4' assegnati alla domanda 1 nelle sessione di settembre 2016 e febbraio 2017.

Per quanto concerne, invece, la domanda 2, possiamo evidenziare che anche in questo caso le valutazioni espresse dai candidati sono state alte, prevalgono infatti i valori '3' e '4' e, come per la domanda 1, nel 2017 sono stati indicati valori più alti rispetto al 2016. Tuttavia le percentuali relative al valore '4' si sono abbassate. Nel 2016, infatti, i candidati hanno prevalentemente espresso, in relazione a tutti gli aspetti, il valore '3' e anche nel 2017, pur rimanendo il valore '4' quello indicato per la maggior parte degli aspetti, le percentuali risultano più basse. L'unico aspetto con la percentuale più alta per il valore '4', sia nel 2016 che nel 2017, è stato ‘prestare attenzione alla formulazione delle istruzioni’ (33% nel 2016 e 41% nel 2017)[4]. Se si considerano, infine, le percentuali relative ai valori ‘1’ e ‘2’, l’aspetto per il quale la somma di tali percentuali risulta più alta (11% e 25% nel 2016 e 10% e 21% nel 2017) è ‘riflettere sul ruolo delle tecnologie e di Internet nell’insegnamento di una L2’.

Si osservino a tal proposito i grafici 3 e 4 e la tabella 2 di seguito in cui sono state riportate le percentuali relative ai valori ‘3’ e ‘4’.

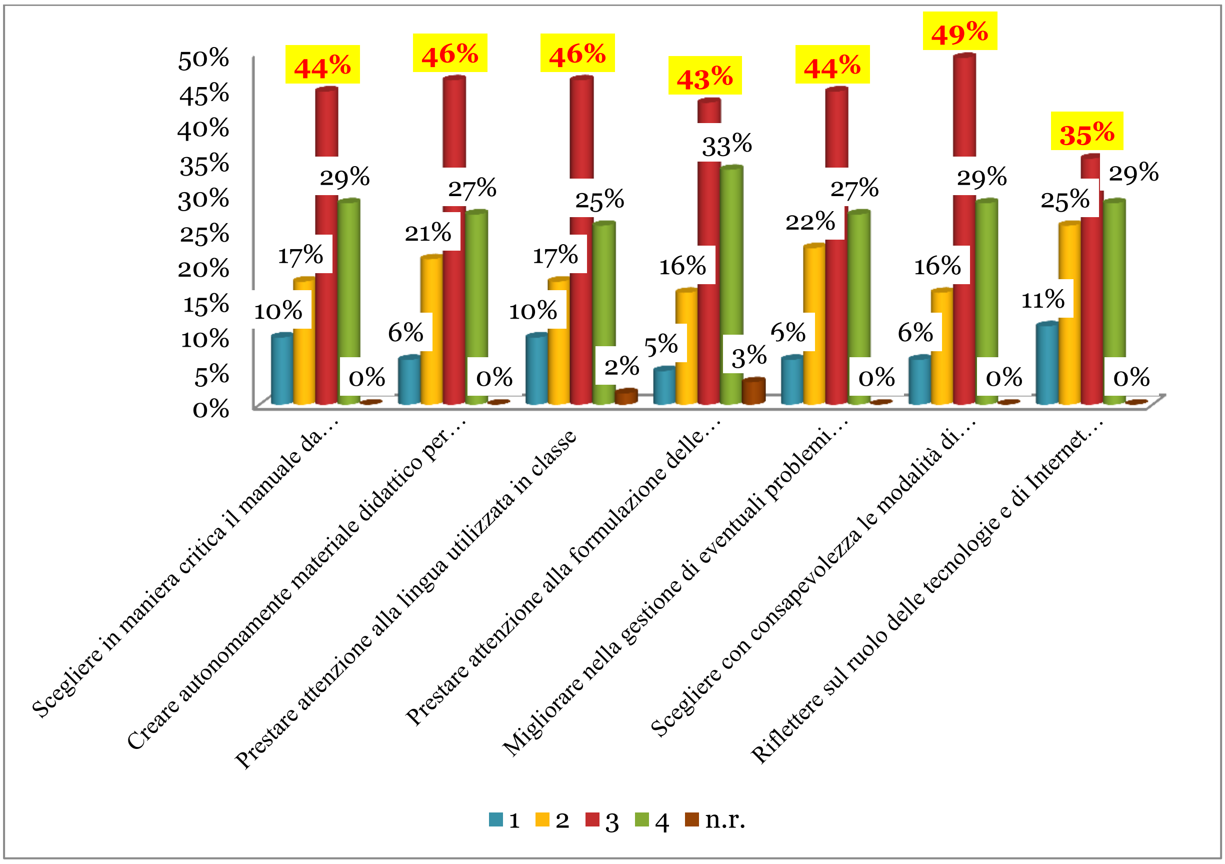


**Grafico 3.** Valori attribuiti dai candidati della sessione di settembre 2016 agli aspetti indicati nella domanda 2 (Dopo essersi preparato/a all’esame DILS-PG di II livello quanto pensa di riconsiderare il Suo modo di insegnare in relazione ai seguenti aspetti? Dia una valutazione da 1 a 4).

[4] Questo dato è significativo in quando dimostra che l’attenzione dedicata all’analisi e alla formulazione delle istruzioni, sia nella prova d’esame che nei corsi di preparazione ad essa, è stata avvertita e riconosciuta dai candidati.

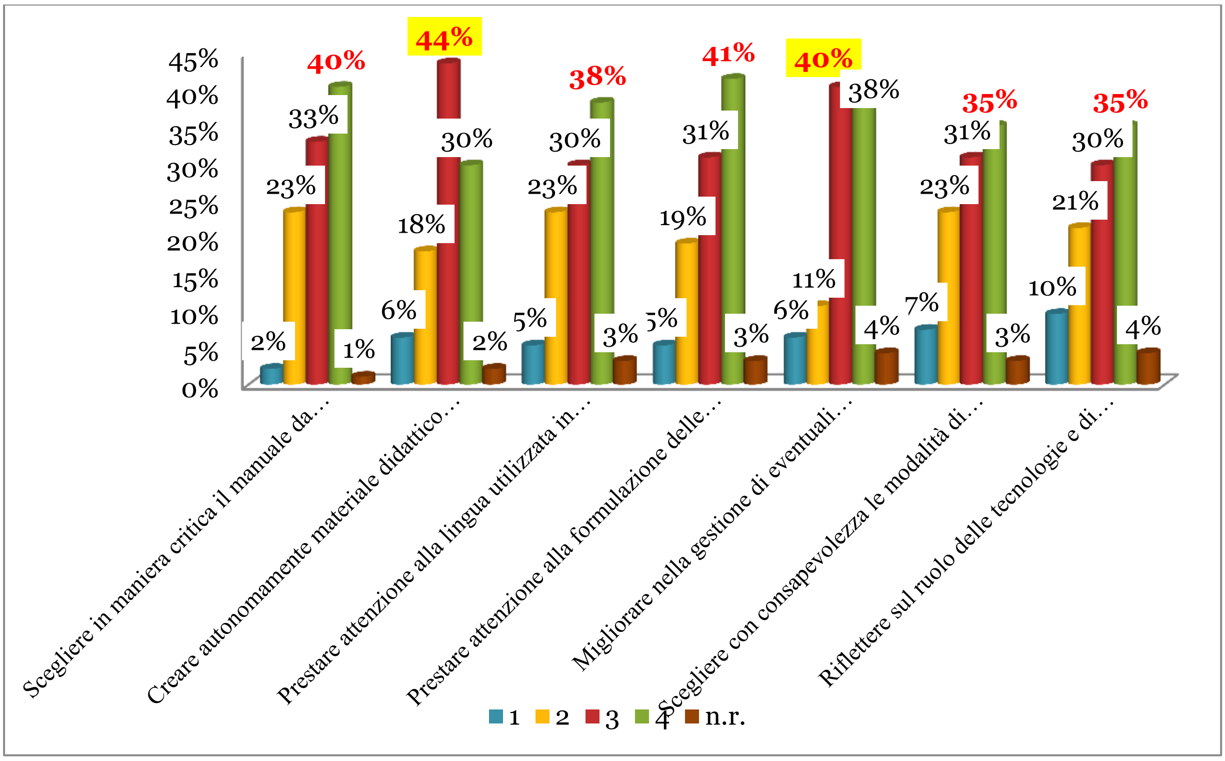


**Grafico 4**: Valori attribuiti dai candidati della sessione di febbraio 2017 agli aspetti indicati nella domanda 2 (Dopo essersi preparato/a all'esame DILS-PG di II livello quanto pensa di riconsiderare il Suo modo di insegnare in relazione ai seguenti aspetti? Dia una valutazione da 1 a 4).

| | 2016 | | 2017 | |
|---|---|---|---|---|
| | 3 | 4 | 3 | 4 |
| Scegliere in maniera critica il manuale da adottare in classe | **44%** | 29% | 33% | **40%** |
| Creare autonomamente materiale didattico per gli studenti | **46%** | 27% | **44%** | 30% |
| Prestare attenzione alla lingua utilizzata in classe | **46%** | 25% | 30% | **38%** |
| Prestare attenzione alla formulazione delle istruzioni | **43%** | 33% | 31% | **41%** |
| Migliorare nella gestione di eventuali problemi che si possono verificare in classe favorendo anche la partecipazione attiva degli studenti | **44%** | 27% | **40%** | 38% |
| Scegliere con consapevolezza le modalità di correzione degli errori commessi dagli studenti in relazione alle circostanze e/o alle finalità dell'attività svolta | **49%** | 29% | 31% | **35%** |
| Riflettere sul ruolo delle tecnologie e di Internet nell'insegnamento di una L2 | **35%** | 29% | 30% | **35%** |

**Tabella 2.** I valori '3' e '4' assegnati alla domanda 2 nelle sessione di settembre 2016 e febbraio 2017.

## 5 Conclusioni

Riepilogando, per quanto riguarda la domanda 1 relativa alla percezione complessiva dell'esame come occasione di sistematizzazione, approfondimento, riflessione e acquisizione di maggiore consapevolezza, stimolo per il proprio lavoro, i candidati si sono espressi positivamente sia nel 2016 che nel 2017, valutando per lo più i singoli aspetti menzionati nel questionario con il valore '4'. Solo per quanto riguarda la revisione del proprio modo di insegnare è prevalsa sia nel 2016 che nel 2017 la percentuale di coloro che hanno assegnato un valore '3' ed è aumentata di conseguenza la percentuale di coloro che hanno attribuito a tale aspetto i valori ‘1’ e ‘2’. Se da un lato l’esame offre l’occasione di riflettere, aggiornarsi, avere nuovi stimoli, ecc., dall’altro è forse più difficile che porti i docenti a modificare del tutto l’impostazione relativa al proprio modo di insegnare, anche se la maggior parte dei candidati si è comunque espressa con il punteggio ‘3’.

Per quanto concerne, invece, la domanda 2 relativa ai singoli aspetti verificati all’interno della prova d’esame, ovvero l’analisi e la creazione di materiali didattici, la gestione della classe e l’utilizzo delle tecnologie e di Internet nell’insegnamento di una lingua straniera, anche in questo caso sono prevalsi i valori ‘3’ e ‘4’, tuttavia c’è stata una prevalenza del valore ‘3’ nel 2016 per tutti i singoli aspetti indicati nel questionario e un abbassamento delle percentuali relative al valore ‘4’ nel 2017, se confrontate con le percentuali rilevate per il valore ‘4’ relative alla domanda 1 (spesso superiori al 50%).

Sembra, quindi, che i docenti del campione di riferimento analizzato riescano a considerare l’esame, nel suo insieme, come occasione di crescita, e ne riconoscano in misura minore l’importanza per quanto riguarda le singole competenze richieste a un insegnante di italiano a stranieri per poter svolgere in maniera efficace il proprio lavoro.

Se si mettono in relazione i dati emersi dall’analisi delle risposte fornite nel questionario e i dati relativi al profilo dei candidati, si può notare che, nonostante nel 2017 i candidati risultino più formati e aggiornati in didattica dell’italiano a stranieri o di altre lingue straniere rispetto al 2016, tuttavia essi considerano l’esame di certificazione glottodidattica DILS-PG di II livello come un’occasione di ulteriore formazione e aggiornamento e, di conseguenza, di ulteriore crescita. Tra i motivi che li hanno spinti a voler conseguire una certificazione glottodidattica come la DILS-PG di II livello al primo posto risulta sia nel 2016 che nel 2017 il desiderio di  crescere come insegnante, motivazione che nel 2016 ha avuto la stessa percentuale di quella relativa alla mancanza di un titolo specifico.

Dai dati emerge, inoltre, che la percentuale più alta di coloro che hanno sostenuto tale esame nel 2016 e nel 2017 è costituita da docenti della scuola pubblica o paritaria che probabilmente hanno sviluppato una maggiore sensibilità e consapevolezza di un bisogno formativo, considerata la presenza di alunni stranieri nelle scuole italiane, oltre al bisogno di conseguire un titolo specifico per l’insegnamento dell’italiano lingua non materna data l’opportunità rappresentata dalla nuova classe di concorso A23. Hanno probabilmente contribuito a ciò anche le recenti iniziative del Ministero dell’Istruzione, dell’Università e della Ricerca (MIUR), come ad esempio il Piano per la formazione dei docenti e la Carta Docenti, che promuovono, tra i docenti della scuola pubblica, una formazione rinnovata e uno sviluppo professionale continuo “per il miglioramento e per l’innovazione del sistema educativo italiano” (http://www.istruzione.it/allegati/2016/Piano_Formazione_3ott.pdf).

Per concludere, i dati confermano che la verifica delle conoscenze e competenze richieste a un docente di italiano a stranieri per poter svolgere al meglio il proprio lavoro può non essere  finalizzata solo alla valutazione, ma costituisce un’occasione di apprendimento, sistematizzazione, riflessione sulla propria esperienza di insegnamento e uno stimolo per un aggiornamento continuo.

## Bibliografia

Ciliberti, A. (2007). Formazione di base e formazione specialistica per l'insegnamento dell'italiano lingua non materna. In E. Jafrancesco (a cura di), *La formazione degli insegnanti di italiano L2: ruolo e competenze nella classe di lingua* (Atti del XV Convegno nazionale ILSA, Firenze 10-11 novembre 2006) (pp.19–34). Roma: Edilingua.

Diadori, P. (a cura di). (2010). *Formazione Qualità Certificazione per la didattica delle lingue moderne in Europa*. Firenze: Le Monnier.

Diadori, P. (2010). La formazione dei docenti di italiano L2: lo stato dell'arte. In P. Diadori (a cura di), *Formazione Qualità Certificazione per la didattica delle lingue moderne in Europa* (pp. 80–93). Firenze: Le Monnier.

Jafrancesco, E. (a cura di). (2007). *La formazione degli insegnanti di italiano L2: ruolo e competenze nella classe di lingua* (Atti del XV Convegno nazionale ILSA, Firenze 10–11 novembre 2006). Roma: Edilingua.

Marasco, M. V., Santeusanio, N. (2016). La consapevolezza metalinguistica dei docenti di italiano L2 negli esami di certificazione glottodidattica DILS-PG. In A. De Marco (a cura di), *Lingua al plurale: la formazione degli insegnanti* (Atti del III Congresso della Società Italiana di Didattica delle Lingue e Linguistica Educativa – DILLE, Rende 8-10 ottobre 2015) (pp. 103–115). Perugia: Guerra.

Rossner, R. (2010). Una griglia di descrittori per il profilo professionale dei docenti di lingue: uno strumento per lo sviluppo professionale e per la gestione della formazione permanente. In P. Diadori, Pierangela (a cura di), *Formazione Qualità Certificazione per la didattica delle lingue moderne in Europa* (pp. 38–47). Firenze: Le Monnier.

Santeusanio, N. (2013a). La certificazione in Didattica dell'Italiano Lingua Straniera: DILS-PG. *AggiornaMenti* (*Rivista dell'Associazione dei docenti di italiano in Germania*) 4/13, 66-78. Retrieved from http://test.adi-germania.org/it/aggiornamenti-04/

Santeusanio, N. (2013b). La formazione all'Università per Stranieri di Perugia. In A. Benucci, Antonella (a cura di), *Formazione e pratiche didattiche in italiano L2* (pp. 221–228). Perugia: Ol3..

Santeusanio, N. (2014a). *Prepararsi alla DILS-PG*. Torino: Loescher Editore.

Santeusanio, N. (2014b). DILS-PG. *Certificazione in Didattica dell'Italiano Lingua Straniera, II livello. Sillabo e linee guida*. Torino: Loescher Editore. Retrieved from http://www.loescher.it/imparosulweb/9788858316603/prepararsi-alla-dils-pg#

Santeusanio, N. (*in corso di stampa* a). Le capacità gestionali dei docenti di italiano L2 negli esami di certificazione glottodidattica DILS-PG. In M. Borreguero, Margarita (a cura di), *Acquisizione e didattica dell'italiano* (Atti del XIV Congresso SILFI Madrid 4-6 aprile 2016).

Santeusanio, N. (*in corso di stampa* b). La certificazione glottodidattica DILS-PG: saperi testati e profilo degli iscritti all'esame di II livello. In C. M. Coonan (a cura di), *La didattica delle lingue nel nuovo millennio: le sfide dell'internazionalizzazione* (Atti del III Congresso della Società Italiana di Didattica delle Lingue e Linguistica Educativa – DILLE, Venezia 2-4 febbraio 2017).

## Sitografia

DILS-PG: https://www.unistrapg.it/it/node/1782

EPG: http://egrid.epg-project.eu/it/egrid (versione in italiano)

Piano per la formazione dei docenti 2016-2019: http://www.istruzione.it/allegati/2016/Piano_Formazione_3ott.pdf

# Certification of Proficiency in Polish as a foreign language and its influence over the Polish labour market

**Dominika Bartosik**, Uniwersytet Jagielloński, Polska

**Abstract:** The main aim of this paper was to answer the following questions: If and to what extent companies in Poland refer to language certificates? What was the level of the required language competences verified by employers? Is knowledge of English enough in the foreign labour market? What is the status of the Polish language? To answer that, the level of foreign language competences and their status among foreigners working in Poland were examined. An important element was also the analysis of the Polish labour market's attractiveness. The results have shown that employers do not refer to the Common European Framwework of Reference for languages (CEFR), but verify the level using other methods described in this paper. Even though English is perceived as a *lingua franca* in business communication, competence in English alone will not suffice for success in the labour market – knowledge of other languages needs to be proved. Research has shown that the status of Polish is improving and becoming increasingly significant in the European context.

## 1 Introduction

The system of certification in Poland is quite new if we compare it to other European countries. Up until 2016, only three levels of exams were available; now there are all levels mentioned in the Common European Framework of Reference for languages (CEFR). This upward trend shows that the importance of a certificate of Polish as a foreign language is growing and this field needs to be explored.

I took into consideration the relation between:

| | | |
|---|---|---|
| Employers | vs. | Employees |
| International companies | vs. | Non- Polish speakers |

The main purpose of my research project was to examine the status of languages in international companies operating in Poland, especially Polish.

This paper fits within the strand “in a globalised economy”. But what does *globalisation* actually mean? We have all heard the term, but mostly connected with fields like international integration, transportation, spreading culture, migration or international commerce. Surprisingly, it is not commonly associated with something which is called *language economics.*

*Language economics* is making use of the methodology of economics to study language in the context of management and the labour market. One thing has to be said and highlighted: the knowledge of foreign language(s) is perceived as a capital investment which is supposed to yield a profit (not only in the financial sense, but some benefits such as developing soft skills in a social context or increased international mobility).

## 2 Literature review

Analysis of this issue required access to English sources, because in Polish literature this is only discussed in the context of other issues. As previously stated, language is perceived as a capital worth investing in, because it is a catalyst of different types of profits. Considering the linguistic diversity not only in Europe but also worldwide, there must be a common way to communicate between countries. This must be taken into account when planning a business strategy (Chiswick, 2008). Nowadays, knowledge of foreign languages is a desirable element, which influences the gross domestic product and economic position of the state (European Commission, 2011). *Language management strategies* is defined as the “planned adoption of a range of techniques to facilitate effective communication with clients and suppliers abroad” (European Commission, 2011, p. 4). From this definition, it can be inferred that the key step to success in language management is to plan such actions and spread them over time, because this guarantees a better placement of the capital. The power of a language is determined by, among others, the number of people who speak it, how widespread it is, and cultural factors (Pawlowski, 2008). In light of this, it is evident that some languages are playing a more important role than others in the business context, and employers are willing to invest in them.

It is widely believed that English is the *lingua franca* in the business world, and that it is enough to use this language to succeed. According to the European Commission (2015), advanced knowledge of English is perceived as a basic skill rather than an asset, as it used to be a few years ago. Nowadays, knowledge of a second language at least at intermediate level is required and this is evidence of spreading multilingualism. Foreign languages skills are definitely a career driver, but only if they are connected to other factors, such as soft skills (European Commission, 2015). Ability to negotiate is an example of a soft skill, while knowledge of the language is a hard skill. Employers require evidence of the desired skills, and certification is one of the methods to verify them. Unfortunately, according to the data of the European Commission (2015), just 1% of employers ask for it during the recruitment process, and some of them do not even know about it. This is the data regarding Europe, so this paper will turn its attention to Poland. How do employers verify the language level? Do they use the CEFR? To what extent do foreigners perceive Poland as an attractive country?

## 3 Methods

### 3.1 Participants

My research study was carried out in international companies operating in many different cities in Poland, such as Kraków, Warszawa, Wrocław, Katowice, Gdańsk, Rzeszów, Poznań and Łódź. This allowed me to gather more representative results and formulate general conclusions. The survey participants comprised 163 individuals from 47 countries outside Poland, working mostly in IT, HR, marketing and finance. 40% of respondents work in the IT sector.

#### *3.1.1 Country of origin*

In total, there were respondents from 47 countries, not only from European countries, but from all around the world. The diverse backgrounds of the respondents was surprising; I personally expected that there would be many more respondents from our eastern border, especially Ukraine (as suggested in all the reports and statistics that I mentioned in the literature review). In fact, they are the most numerous group, but there are not many more of them than respondents from Spain, Italy or France. We can suppose that the reason for this is  perhaps the industry of business they are working in, and the difficulty with finding employment in their own countries, as suggested by other reports by the European Commission (2017)

### *3.1.2 Industry of  the business*

There was a significant number of respondents working in the IT industry, as mentioned previously. Then there were many people working in the field of IT, HR, marketing and finance. Other fields were also mentioned, but the number of occurrences was low.

When asked about their job title, repondents provided a wide range of answers: web developer (42%), project manager (21%),  business analyst (15%), software developer (12%), HR administrator (6%) and other (4%).

IT-related jobs use programs (Java, PHP, Python etc), which are all written in English, so initially they don't need any other language except English to do their jobs. This language is required by employers.

### *3.1.3 Years of experience*

Most of the respondents answered less than 1 year. The shortest length of time was 9 months and the longest was 30 years. The average was 5.5 years.

## 3.2 Instrument

The questionnaire consisted of 20 questions: 16 were closed, 4 open. Before completing the questionnaire, respondents were asked to fill in some basic information to capture a more rounded profile of the participants. Filling in the online survey took 3-5 minutes, depending on how detailed their answers were.

## 3.3 Data collection

To collect answers I used a website program[5], which also analyses data and suggests patterns in the answers.

## 3.4 Data analysis

The most important element was the collection of a sufficient number of questionnaires to formulate conclusions. The first stage was the analysis of the first part, which concerned the general linguistic situation of companies with foreign capital. The second part concerned only the situation of the Polish language in the labour market. Observation of the results of these two stages allowed me to compare the status of the Polish language with other languages.

The results I obtained were surprising, but some of them comfirm the assumptions of the European Commission reports, which were mentioned in the previous section.

[5] www.profitest.pl

## 4 Results

I will not analyze in detail all of the answers and questions in the questionnaire. Instead, I have selected what I consider the most relvant to answering the research questions.

(1) Do you agree that nowadays languages are a career driver?

78% of the respondents answered that they consider languages a career driver. According to the European Commision's report (European Commission, 2015), this confirms the assumption that this is true if it is supported by other relevant (specific) skills.

(2) How many languages do foreign workers in Poland know?

The results from this question are really promising, because the average is 2.83 languages. I would like to highlight that the question was about foreign languages, so the mother tongue doesn't count in this ranking. It confirms the effectiveness of the European initiative of spreading plurilingualism among the citizens with the goal of mother tongue +2.

(3) Which languages are considered most important to Polish companies?

Respondents' opinion was that English is the most important language, which is not a surprise, then German and Spanish.

(4) What is the most dominant language in Polish international companies?

Since knowledge of English is practically obligatory, and it is the language of international communication, what is then the status of the Polish language? Polish was mentioned as dominant in the company only by 19% of the respondents, so which languages were required in the recruitment process?

Polish was obligatory in 7 % of answers. In most cases, two languages were required: English, no longer considered an added benefit but a necessary, basic skill; and at least one additional language. Among additional languages, participants mentioned not only well-known and popular languages, but also Hungarian, Bulgarian, Norwegian, etc.

(5) How do companies verify the level of the required languages?

For this question, participants had the possibility of selecting more than one answer, because they are not mutually exclusive. The predominant way of verifying is the interview in the required language. Written and/or oral tests were also used frequently. It is worth noting that certification was mentioned only by 9.1% of the respondents. Nevertheless, if they were asked about their certificates in any language, only 46% confirmed that they have certificates. Since more than half of the respondents do not have a certificate in any foreign language, it can be deduced that certification is not the main way of verifying language skills in Polish companies.

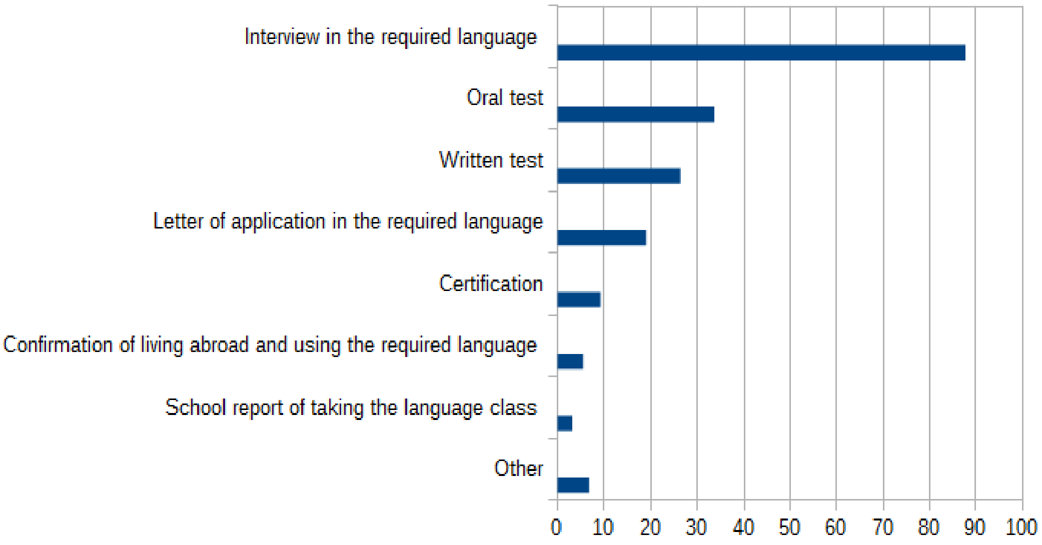


**Figure 1.** How was the level of the required language verified?

(6) How do companies specify the language level required in their recruitment announcements?

As shown in the chart, most of the employers used expressions like: fluent, basic, passive knowledge. A quarter of the respondents mentioned the levels A1–C2 of the CEFR. Some of them did not even mention a level. According to the European Commission, employers may avoid using the CEFR levels because they do not want to discourage people to apply for a particular position (European Commission, 2015).

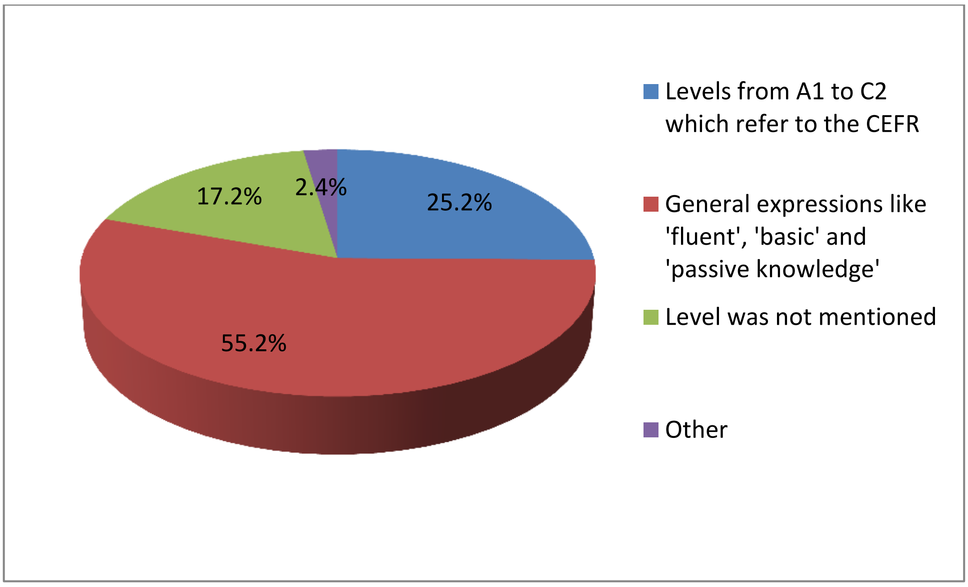


**Figure 2.** Which expressions were/are mentioned in the recruitment announcements of the company regarding foreign languages?

(7) How many foreign employees know Polish, and at what level?

60% of foreigners declare that they know the Polish language. Nearly 26% estimate their level as intermediate, which would be B1–B2 according to the CEFR. 22% of the respondents considered themselves beginners, and 21% declared an advanced level of the Polish language.

They were also asked about their knowledge of Polish before starting work in Poland. Only 40% said that they knew some Polish before moving to Poland. The results also show that 20% of the respondents started studying Polish after moving to this country.

(8) What kind of support do companies offer to enhance the level of proficiency in Polish among their employees?

There are obviously different ways to learn the language. 42.6% of the respondents confirmed that their companies pay for language courses during working hours, almost 15% said that their companies pay for private lessons, some of them provide e-learning, and others have Polish lessons at work, but before or after working hours.In total, only 29.5% of employers have no formal policy regarding the improvement of employees' language skills. Interestingly, by contrast, the European Comission (European Commission, 2015) states that just below half of the companies pay for language course training for their employees. Considering the results from my study, this could also mean that employers do not invest generally in new languages; they prefer to improve the existing ones and invest in improving technical, job-related skills instead.

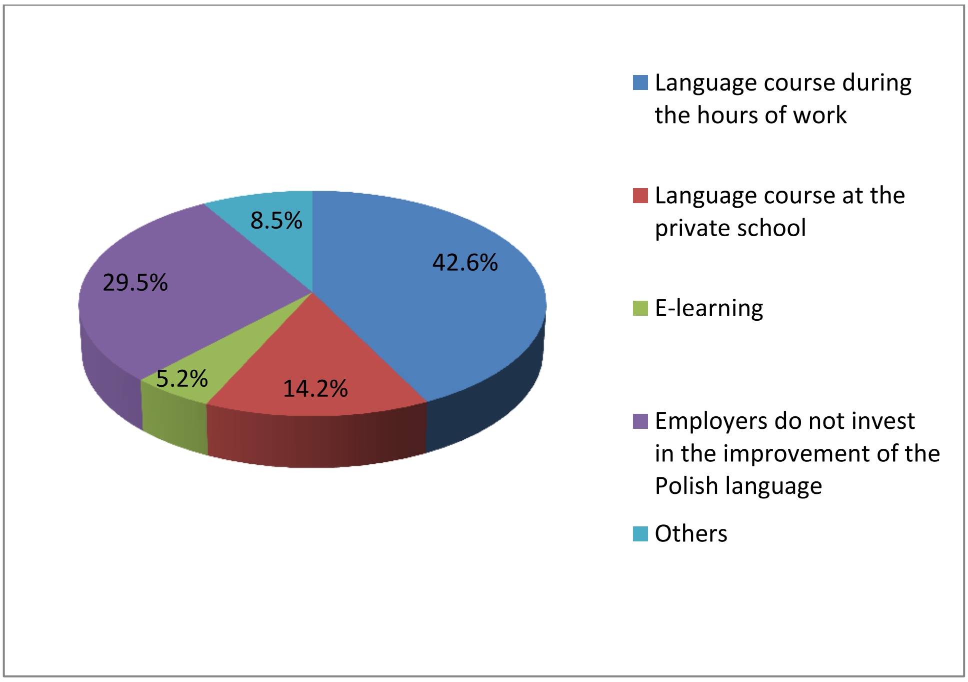


**Figure 3.** What kind of support does the company offer to enhance the level of proficiency in Polish among employees?

(9) Do employees use specialised vocabulary in Polish in their daily work?

One of the aims of this research was to check how the certificate of Polish as a foreign language is perceived among the employers and employees. It is the first step toward a plan of action to raise public awareness about it and to plan the implementation of specific kinds of certificates, such as a business variant. When I asked respondents if they generally use specialised vocabulary in their day-to-day work (without specifying the language), almost 70% answered yes.

One can therefore conclude that it would be advisable to create and implement a business variant certificate to prove linguistic knowledge in the given field. 57% thought that it would not be useful, the rest stated that it could be. This last answer is not surprising, especially if we consider that  60% of the respondents had not even heard about the existence of the general certificate of Polish language.

(10) How do employees rate the status of the Polish language compared to other European languages? (Respondents could only choose one option to this question.)

Half of the respondents answered that Polish is considered as 'other language' (not one of the main languages of the European Union) in the ranking. 21.8% claim that it is not valuable at all, 19% answered that it is becoming more and more significant, and 8.6 % consider it as one of the top 10 languages in Europe. This is a bit surprising because it seems that the position of the Polish language continues to go up – Poland is considered a rapidly developing country (European Commission, 2016). According to the same report mentioned above, the status of the Polish language is likely related to the economic growth and the number of speakers worldwide.

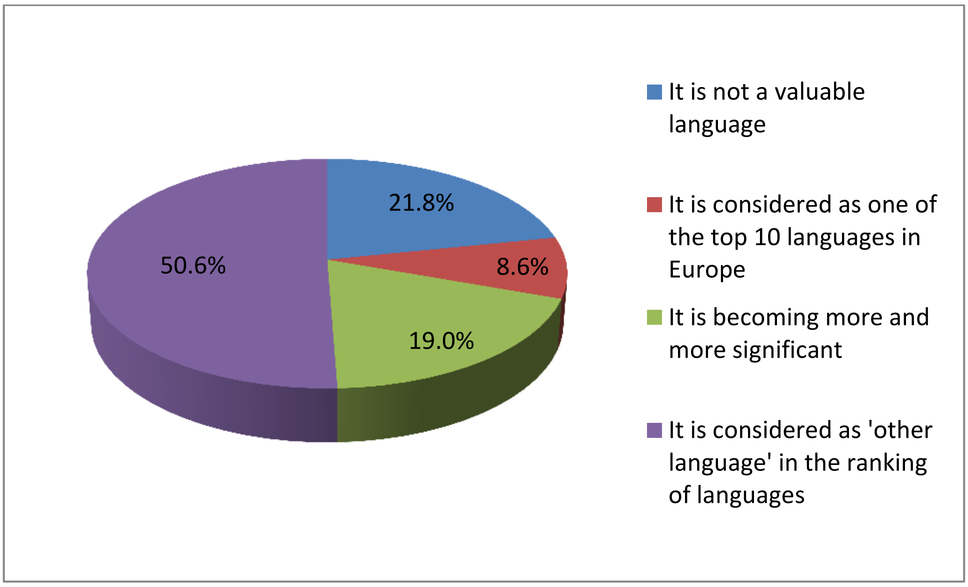


**Figure 4**. How do you rate the status of the Polish language compared to other European languages?

## 5 Conclusion

The importance of the Polish language is growing, and more and more international companies invest in the improvement of it among their foreign employees. Even though English still maintains the dominant position, it is now perceived as a necessity rather than an additional benefit; thus, multilingualism is spreading. Certificates and the formal classification of the levels mentioned in the CEFR levels are not currently a commonly used way to verify the knowledge of language skills.

Poland became a member country of the European Union in May 2004; the same year saw the creation of the certificate of Polish as a foreign language. Since then, both the position of Poland in the European Union and the importance of the certificate have grown significantly, so if the appropriate steps are implemented, we may fulfil the potential this certificate can provide.

## References

Chiswick, B. (2008). *The Economics of Language: An Introduction and Overview*. Retrieved from http://ftp.iza.org/dp3568.pdf.

European Commission. (2011). *Report on Language Management Strategies and Best Practice in European SMEs: The PIMLICO Project*. Retrieved from http://ec.europa.eu/dgs/education_culture/repository/languages/policy/strategic-framework/documents/pimlico-full-report_en.pdf

European Commission. (2015). *Study on Foreign Language Proficiency and Employability. Final Raport*. Retrived from http://www.erasmusplus.sk/kniznica/publikacie/Final_Report.pdf

European Commission. (2016). *2016 European Semester: Country Report – Poland*. Retrived from https://ec.europa.eu/info/publications/2016-european-semester-country-report-poland_en

European Commission. (2017). Joint Employment Report 2017. Retrived from *https://ec.europa.eu/info/publications/2017-european-semester-draft-joint-employment-report_en*

Pawłowski, A. (2008). Zadania polskiej polityki językowej w Unii Europejskiej. W. In J. Warchala & D. Krzyżyk (Eds.), *Polska polityka językowa w Unii Europejskiej* (pp. 113–147). Katowice: Wydawnictwo Uniwersytetu Śląskiego.

# Beliefs Driving the Assessment of Speaking: An Empirical Study in a Brazilian Public Classroom

**Eber Clayton Dutra**, University of Brasilia, Brazil
**Gladys Quevedo-Camargo**, University of Brasilia, Brazil

**Abstract:** In the Brazilian educational system, the development of the students' speaking ability has traditionally been left aside for a number of reasons, except in some very specific school contexts. Considering the importance of speaking, the crucial role assessment plays in the teaching–learning process of all language abilities, and the relevance of language assessment literacy in teachers' professional development (Fulcher, 2012; Scaramucci, 2016; Taylor, 2009), this on-going empirical study aims at investigating how the beliefs (Brown, 2008; Pajares, 1992), especially beliefs about oral fluency (Chambers, 1997; Koponen & Riggenbach, 2000), held by English language teachers of a particular public school drive how their students' speaking production is assessed. Partial data analysis of this qualitative research has shown that five of these detected beliefs affect, in different ways, classroom practice, some quality principles of the oral tests and, consequently, the outcomes.

## 1 Introduction

In Brazil, just like in the international scenario, speaking foreign languages has been a growing demand. Internationally, factors such as the processes of globalisation, thus making global communication easier as well as the economic crises, environmental catastrophes and wars have triggered migration movements along with the need to learn foreign languages for professional and academic purposes (Fulcher, 2012; Scaramucci, 2016). Nowadays, in the Brazilian context, besides allowing access to cultural possibilities, knowing a foreign language is an important element both for academic education and qualification for the labour market. In this sense, speaking a foreign language fluently can enhance employability and increase salaries, as well as facilitate career advancement.

As Scaramucci (2016) shows, if there is demand for proven knowledge in foreign languages, likewise there is demand for assessment competences specific to the assessment in language contexts. As a consequence, general, academic and specific purposes proficiency exams abound, thus causing proportional growth in the importance of testing and assessment, certification processes and discussions about practices and quality standards to be adopted.

Such quality standards nourish the ideas in the current study. Questions were inspired from our pedagogical practice as EFL teachers in public language schools in Brazil, as well as from several issues raised by our students. Among such issues, "oral fluency" stood out as a main concern of reflections on learning assessment. Do teachers/assessors have the same concept of fluency when they assess their students' oral production? What do they believe oral fluency is? The search for answers led us to the following research question: How do the beliefs about oral fluency held by Brazilian English language teachers influence the way their students' speaking production is assessed? After all, as Guillot (1999) states, fluency is a vague concept whose justification is rarely based on anything but intuition.

Studying teachers' intuitions and conceptions of oral fluency led us to the study of beliefs – and the relationship between beliefs and assessment. A conception is a mental construct or representation of reality that contains beliefs, concepts, preferences and meanings, among others (Thompson, 1992), and, according to Brown, Lake, and Matters (2011), there is solid evidence indicating that teachers' conceptions of how content is learned, taught and assessed influence the way they teach and what students learn. Beliefs, in turn, are referred to by means of a wide variety of terms in the literature (Pajares, 1992, p. 309), such as "values", "judgements", "opinions", "perceptions" and "personal theories". According to this author, belief structures or systems designate the set of beliefs individuals, individually or collectively, hold about a particular topic.

## 2 Method, context and participants

This article reports on an on-going qualitative interpretivist and ethnographic study whose focus is on the oral assessment process carried out by a group of EFL teachers, specifically in respect of their beliefs on oral fluency.

According to Erickson (1984), ethnography is the study of the culture of a social group. This particular group of teachers works in a public language centre in the city of Brasília, Federal District (the country's capital city). The data were collected from August to November 2016.

Aiming at offering students from public schools optional and additional higher quality studies on foreign languages (Spanish, French, English and Japanese), the public language centres started operating in 1975. Nowadays there are 15 units functioning in the 14 administrative areas of the Federal District. The centre where the data were collected has 12 teachers of English, five teachers of Spanish and three teachers of French, and offers courses to 3000 students in the morning, afternoon and evening – 450 of them in the evening only.

The participants of the research are four English teachers aged 28 to 45 who teach in the evening and two coordinators, all of them majored in English language (Letters course), including the coordinators, with no post-graduate courses or international certificates. They learned English at English language schools in Brazil, and have worked as English teachers for over three years (one of the participants), 10–15 years (two participants) and 16–20 years (three participants).

Classroom ethnography entails intense and detailed classroom observation, usually for a term or a year depending on the school period, complemented by audio or video recordings of activities, interviews with teachers and students, diaries, as well as descriptions of the school and classroom environments. It enables the researcher to examine, among other issues, teaching styles and approaches, lesson structure, teachers' and students' discourses and their expectations, idealisations and resistance. Thus, the instruments and techniques used to collect and select the data for this study were a questionnaire (to obtain information on the participants' professional and academic background, for instance), a semi-structured interview, classroom observation, field notes and audio recordings of the interviews and oral assessments.

In order to help answer the research question, some of the questions asked in the interview were: For you, what does knowing a foreign language (English) mean?; What do you like working on best in your lessons: writing, grammar, vocabulary or speaking? Why?; If you had to choose a general objective for your lessons, what would that be?; How do you define 'fluency'?; How do you identify a non-fluent speaker of English?; Is it possible to teach fluency?; In your point of view, what is the best way to assess an EFL student's oral fluency?

The instruments and techniques chosen for this study proved to be appropriate for data collection, which facilitated the next step of the research, data analysis, discussed in the following section. According to Bogdan and Bicklen (1998), data analysis is the process of research and systematic organisation of observation field notes, interview transcripts, and other rough materials the researcher accumulates to study. It involves, in addition to coordinating, synthesising and categorising the data, finding out what is relevant and deciding what will be said in the report – or ethnography.

## 3 Preliminary results and discussion

In this first stage of the research, the interviews revealed a variety of beliefs in relation to oral assessment. Among them, five beliefs on oral fluency the teachers have in common are highlighted below.

(1) Belief 1: As English teachers, we share a common definition of fluency.

The first belief identified is related to the fact that all the four interviewees believe they share the same concept of fluency as their workmates'. They all reported never having talked about or reflected upon the concept or the nature of the fluency they were assessing. Though different viewpoints coexist, there is no reflection on the fluency construct. This apparent consensus is a mistake as the same student, when assessed by two or three teachers, would be assigned different marks. With no clear and defined criteria, the test validity (Bachman, 1990) can be questioned: what is being measured?

(2) Belief 2: Speed, smoothness and effortlessness are the main features of a fluent performance.

This historical meaning of fluency as flow, harmony and continuous stream, involving temporal, acoustic and phonetic characteristics of speech is sustained not only by laypeople, in the non-technical use of the word, but also by EFL professionals (Koponen & Riggenbach, 2000). In this perspective, fluency is mainly connected with speed and lack of excessive pauses in speech – criteria generally used as a reference to assess fluency in speaking tests. However, Riggenbach (1991) shows that neither speed nor low frequency of pauses would guarantee that a speaker is considered fluent. To reduce fluency to aspects such as speed and lack of hesitation constitutes a simplistic view as fluency requires several other requisites as, for instance, the role of comprehension lapses. Another aspect to be analysed is that the descriptors used for fluency are not linguistically specific. In this study, the interviewed teachers could not reach consensus as to what "fluid", "smooth", "fragmented" or "hesitant" are.

Another problem adjacent to this belief lies in the relation between this school's teaching and assessment approaches and the teachers' assessment practices. The school's English course is labelled as communicative, aiming at (and assessing) effective communication by means of students' debates, interviews and presentations. The fluent language, including reading and writing, can be accurate, but that would not be the student's focus, nor the focus of the teaching offered by the school. However, the observation of the teachers' lessons shows an aspect that is even more traditional in Brazilian language classrooms: the strong emphasis on accuracy, on the study of grammar, on the domain of the language structures, on the focus on form rather than on language use (Brumfit, 1984), which causes lack of alignment with the assessment policy implemented in this language centre.

(3) Belief 3: Practice helps/improves oral production.

The interviewees believe that practicing the language makes fluency development easier, and a list of practice activities was mentioned by the teachers. However, they were not sure about how that development process occurs. As it is not clear how it happens, there is no agreement on what kind of activities can work best for oral fluency. As classroom observation confirmed, written and oral grammar tasks dominated teaching, with rare exceptions, despite the communicative format of the oral tests at the end of the term. Therefore, oral assessment ends up testing an ability that was neither taught nor practised. The test validity can be questioned again: it is fair and appropriate to test what we teach and what the students learn. Also, the negative washback has to be considered (Bailey, 2005; Quevedo-Camargo, 2014): a test under those conditions does not inspire or encourage students to prepare for speaking tasks or promote the development of the skills or knowledge to be learned.

(4) Belief 4: Studying at home and living abroad are the best ways to develop oral fluency.

Becoming fluent seems only to depend on the students and their attitudes outside the classroom. Besides living abroad or travelling to countries where the target foreign language is spoken, the best ways to become fluent – mentioned unanimously by the research participants – are reading books and magazines in the language, talking to native speakers (also on the internet), watching films and listening to music. The observed assessor/teacher frequently acted in accordance with this perspective during the English classes: little was done in class toward the development of oral fluency. For different reasons (particularly time and program control), most of the time the focus was on the students' coursebook and its written exercises.

(5) Belief 5: Fluency is a synonym for oral proficiency.

This notion of fluency complements the second belief mentioned previously. The participants presented a common sense definition of fluency: in lay terms, fluency refers to general proficiency in a foreign language and is related to the effective use of the language, with 100% of domain and control. It is a fact that, during the application or scoring of the oral tests, no reference was made to any assessment scale, and the established parameters seemed to be "good", "average" and "bad". As Chambers (1997) writes, fluency is not proficiency but only one of its descriptors, and one of the risks of thinking they are synonyms is the blend with the notion of "native-like speech", an expression commonly used to describe a competent foreign language speaker. In addition, native-like performance is taken as the linguistic model, and accuracy is highly valued.

## 4 Final considerations

Data discussion highlights two important aspects of oral production assessment as well as of the learning assessment in general. The first aspect is the importance of the study of teachers' beliefs so as to better understand what happens in the classroom and, consequently, in the test. The identification of beliefs about oral fluency in this study supports the view that beliefs influence practices and outcomes (Brown, 2008). Therefore, to alter teachers' assessment practices it is crucial, among other things, to change some teachers' beliefs. Although the results discussed here are preliminary, pedagogical meetings and discussions about beliefs, culture of assessment and different concepts of fluency are already in this language centre's agenda to be started in the second semester of 2017.

The second aspect highlighted here is the importance of specific knowledge (and skills) on assessment matters teachers and instructors need to have for good testing practice – or assessment literacy, as discussed in Taylor (2009). That is not a job only for experts: by holding a background or training experience in assessment, teachers and coordinators involved in learning assessment, like the participants in this research, would be more adequately equipped for this role. That is, just as happens to language test developers and researchers or to those involved in large-scale tests produced by professional organisations, language teachers 'need some measure of assessment training if they are engaged in developing, scoring, interpreting, and improving classroom-based assessments' (Taylor, 2009, p. 24). The way the beliefs detected in this study put at risk the quality principles of the assessment of speaking is an example of how the lack of a better understanding of assessment can impact the form in which the students' speaking production is assessed – and, consequently, disfavour both the assessment area and the foreign language teaching and learning in general.

## References

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bailey, K. M. (2005). *Practical English language: Speaking*. New York, NY: McGraw-Hill.

Bogdan, R. C., & Biklen, S. K. (1998). *Qualitative research for education: an introduction to theory and methods*. Needham Heights, MA: Allyn & Bacon.

Brown, G. T. L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students*. New York, NY: Nova Science Publishers.

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education, 27*(1), 210–220.

Brumfit, C. (1984). *Communicative methodology in language teaching: The roles of accuracy and fluency*. Cambridge, UK: Cambridge University Press.

Chambers, F. (1997). What do we mean by fluency? *System, 25*(4), 535–544.

Erickson, F. (1984). What makes school ethnography 'ethnographic'? *Anthropology and Education Quarterly, 15*, 51–66.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

Guillot, M-N. (1999). *Fluency and its teaching*. Clevedon, UK: Multilingual Matters.

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5–24). Ann Arbor, MI: University of Michigan Press.

Pajares, F. M. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*(3), 307–332.

Quevedo-Camargo, G. (2014). Efeito retroativo da avaliação na aprendizagem de línguas estrangeiras: Que fenômeno é esse? In K. B. Mulik & M. S. Retorta (Eds.), *Avaliação no ensino-aprendizagem de línguas estrangeiras: diálogos, pesquisas e reflexões* (pp. 77–93). Campinas, SP: Pontes Editores.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes, 14*(4), 423–441.

Scaramucci, M. V. R. (2016). Letramento em avaliação (em contexto de línguas): Contribuições para a Linguística Aplicada, educação e sociedade. In C. L. Jordão (Ed.), *A Linguística Aplicada no Brasil: Rumos e passagens* (pp. 141–165). Campinas, SP: Pontes Editores.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21–36.

Thompson, A. (1992). Teachers' beliefs and conceptions: A synthesis of the research. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127–146). New York, NY: Macmillan.

# Assessment in a Globalized Economy: A Task-based Approach to Assess the Proficiency of Dutch in Specific Occupational Domains

**Sarah Smirnow**, CNaVT, Belgium
**Christina Maes**, CNaVT, Belgium
**Lucia Luyten**, CNaVT, Belgium
**Steven Verheyen**, CNaVT, Belgium

**Abstract:** For language tests to be in tune with the target context, a constant attention to the shifting characteristics of real-world language use is required. This holds particularly true for language tests in the occupational domain since domestic labor market demands are continuously changing and increasingly met through the recruitment of foreign workers. This paper describes how subject experts were involved in the cyclical validation process of a test of Dutch for the professional domain. A survey of recruitment agents, employers, policy makers, language instructors, examiners, and former test takers indicated that a task-based test targeting the language skills involved in service-oriented work settings such as administration and health care at level B2 of the CEFR was favored. The involvement of subject specialists recruited among the test's various stakeholders proved to be of vital importance throughout the development and validation process to ensure content validity and avoid biases.

## 1 Introduction

Tests of language for specific purposes (LSP) assess context-specific language performance (Douglas, 2001). The methods and material of LSP tests therefore need to be informed by an analysis of the target language use situations (Douglas, 2000). This poses a particular challenge for developers of occupational language tests since in a globalized economy the formal and informal language requirements for successfully navigating the workplace are subject to constant change. Adapting tests to these changes is a constant concern for test makers as they seek to ensure the validity of their tests. In order to include the relevant content knowledge in the resulting test a strong collaboration with stakeholders is required throughout its development (ALTE, forthcoming). This paper describes how stakeholders were involved in the development of a test of Dutch for the professional domain (PROF) by the Certificate Dutch as a Foreign Language (CNaVT, www.cnavt.org) and the particular challenges it posed.

## 2 Background

The CNaVT is a project of the Dutch Language Union, an intergovernmental organization that promotes the learning and use of the Dutch language across the globe (www.taalunie.org). Dutch and Belgian employees based at the Centre for Language and Education (KU Leuven, Belgium) collaborate with specialists and organizations in both the Netherlands and Belgium to develop task-based (Van Gorp & Deygers, 2013) and domain-specific (Gysen & van Avermaet, 2005) exams of Dutch as a foreign language. All exams are related to the Common European Framework of Reference (CEFR) and are especially developed for higher educated (young) adults who want to prove their proficiency in Dutch as a foreign language with an internationally recognized certificate. The paper-based CNaVT exams typically test the key language skills in an integrated manner (Cumming, 2013) where appropriate (i.e., more integrated testing at higher levels of language proficiency).

In this paper we will focus on PROF, a broad LSP test targeted at the general professional domain (as opposed to narrow LSP tests targeting specific occupations; ALTE, forthcoming). This test was recently renewed, giving the CEFR a more central role in the design of the construct, the task specifications, and the rating model. As part of a cyclical test validation process, the target audience and their needs were also surveyed. In Dutch speaking companies in Belgium and the Netherlands approximately 80 percent of the daily communication is done in Dutch (van der Meulen et al., 2016), but there are few official regulations regarding the language skills required to work in settings where Dutch is the main language of communication (we are only aware of the B2 requirement for doctors, dentists, pharmacists, psychotherapists, and health care psychologists to become accredited in the Netherlands). From the onset it was therefore clear that subject experts were to be involved to get a better understanding of what the test was supposed to assess.

## 3 Test development

### 3.1 Involvement of subject experts

Both the CNaVT test developers and the CNaVT advisory board, made up of assessment experts and teachers of Dutch as a foreign language, were involved throughout the development process of PROF.

Subject experts were recruited from among the different stakeholders at the start of the development process. An invitation for help was extended to a large number of stakeholders, who were also asked for referrals to other relevant actors in the field.

(1) Private recruiters and international mobility managers of the Dutch and Belgian public employment services were contacted since they are familiar with the national labor market demands and therefore ideally placed to identify the jobs foreign people are recruited for.

(2) Domestic employers, language policy makers, and teachers of Dutch as a foreign and/or second language were contacted for the same reason.

(3) CNaVT examiners from all over the globe were involved to help identify foreign companies where employees need to be proficient in Dutch.

(4) Finally, test takers who had passed the former LSP test and were working in jobs that require Dutch, were contacted to get a better insight into the language use and language tasks they were actually performing in Dutch.

### 3.2 Needs analysis

A needs analysis was carried out among the convenience sample consisting of recruitment agents, employers, policy makers, language instructors, examiners, and former test takers. The stakeholders were surveyed about the profile of the working professionals who could benefit from a test of Dutch for the professional domain, the domains and topics appropriate to include in the test, and the required language skills and performance standards.

The subject specialists who helped us to determine the target audience and their language needs confirmed that the occupations targeted in the original test were still very relevant in 2017. That is, they indicated that future test takers were likely to end up working in administrative and/or service oriented jobs, of the kind found in banks, embassies, call centers, and (international) companies dedicated to export or import. A new finding was that recruiters in the Netherlands and Belgium are increasingly turning abroad for jobs in health care, recruiting mainly foreign dentists (Netherlands) and nurses (Belgium).

The initial findings from the needs analysis were subsequently confirmed by a purposive sample comprised of subject experts from the identified occupation domains (including health care professionals such as nurses, care givers, and directors of hospitals and assisted-living centers) and by an independent literature review (e.g., van der Meulen et al., 2016).

### 3.3 Test and task construction

Based on the findings of the needs analysis, the test's original target group was extended to include people working in health care in addition to administrative/service oriented professionals. A purposive sample of subject experts from these domains was put together to aid in the development of the test construct and the ensuing tasks (ALTE, forthcoming; Douglas, 2000). The purposive sample was consulted throughout the development of the test construct and the ensuing tasks. They provided feedback on the relevance and authenticity of the solicited communicative acts, on the tasks' ability to establish whether a test taker masters the necessary language skills to a sufficient degree, and on the susceptibility of the test construct and tasks to cultural biases.

The subject experts favored authentic real-life tasks for the test, but also demanded the tasks not be overly specific so as not to exclude potential test takers (see Brunfaut, 2014, for a discussion of practicality vs. specificity). They indicated that the B2 level was a minimum requirement for these occupational profiles, but hastened to add that generally speaking, foreign employees do not meet this requirement when they are recruited. The subject experts therefore intended to administer the test to employees who had worked and lived in a Dutch environment for a while, as an incentive for them to learn the language or to decide about extending or improving their contract.

CEFR experts, who were particularly familiar with the B2 level and/or the occupation domain, judged the level of the resulting tasks and set an appropriate standard. It proved difficult to include the purposive sample of experts in these stages since many of them were not familiar with the CEFR.

### 3.4 Piloting

Because they were involved early on in the development process, a number of language schools, institutions offering Dutch for occupational purposes, and employers such as hospitals were willing to engage in structural partnerships for piloting purposes. Examiners and pilot test takers at these organizations conveyed their ideas about the relevance, authenticity, and difficulty of the test tasks in interviews with the test developers.

It is common procedure to have test takers and examiners provide feedback on the CNaVT test they took/administered. Their feedback will be analyzed together with the final test results and taken into account while developing future instantiations of the test.

### 3.5 Challenges

Working closely together with subject specialists was a prerequisite for the development of a valid test of Dutch for the professional domain, but not without challenges:

(1) We found that many practitioners in the field were not familiar with the CEFR or entertained very different interpretations of the framework than the professional test developers involved.

(2) The absence of official guidelines on Dutch at the workplace made it difficult to convince stakeholders that using a standardized test related to the CEFR has an added value compared to their own (often idiosyncratic) assessment practices.

(3) The demands imposed on the subject experts are quite high, while there is little immediate return for them. There is no guarantee that their investment will pay off in the long term as the requirements in the job market can be quite volatile and what is a requirement now, needn't be a requirement in the not so distant future. This made it difficult to find subject experts who were willing to engage in the development process.

## 4 The resulting test

PROF is a paper-based test of Dutch in the occupational domain developed for learners of Dutch as a foreign language who want to use Dutch in an occupational context, more specifically in health care or administrative services. The test assesses the key language skills involved in varying work settings that are highly service oriented (e.g., customer service, reception, purchasing department, residential care center, hospital) at level B2 of the CEFR. Communication partners can be unknown (customers, new suppliers, etc.) or familiar (colleagues, patients, known suppliers, etc.). There is no subject-specific knowledge of vocabulary required and test takers are allowed to use a dictionary.

The test consists of three parts with two task-based and highly integrated tasks each. The task introductions have a motivational and clarifying character, and are designed to elicit a well determined response from the test taker. They describe the working environment the test taker finds herself in, her role in this environment, and the nature of her task. The introduction is followed by the instructions, which elaborate the task requirements and guide the test taker to a semi-authentic spoken (Part 1) or written input (Parts 2 and 3).

Task types are:

(1) Part 1: writing a text, based on informative or argumentative spoken input, eventually adding an argumentative part;
(2) Part 2: writing a text, based on informative or argumentative written input, eventually adding an argumentative part;
(3) Part 3: holding a formal or informal plea, based on a very short informative input.

The difference between the first and second task type lies in the nature of the input: the former always has spoken input (a monologue such as a voicemail or short lecture, or a dialogue such as an interview), the latter always has written input (all kind of articles, or a part of a document such as a contract, safety regulations, a brochure about a product, service, or workshop, etc.).

Examples of the output for the first and second task type are writing an e-mail to a superior to convince her to introduce a new regulation or ask for a leave, writing an e-mail to a client to communicate a decision, providing clarification about a product, or writing a note to a colleague about something that happened during the time she was absent.

As a response to the third task type, the test taker could be asked to hold a small presentation about a workshop she attended, to introduce a new colleague to the company, or to perform a job interview. The input for this task type can be varied, but is always a short written text (part of an article found in a newspaper or popular scientific magazine, a brochure of a center for adult education, a company’s website, etc.).

## 5 Conclusion

For language tests to be in tune with the target context, a constant attention to the shifting characteristics of real-world language use is required. This might be particularly true for language tests in the occupational domain. Given that there are few official requirements for Dutch language proficiency in the workplace and the labor market demands are quite volatile, a high level of involvement of subject experts was necessary to develop a test of Dutch for the professional domain. This paper describes how subject experts were involved in the various stages of the development process: the needs analyses, test and task construction phase, piloting, relating the test to the CEFR, and the standard setting. While their involvement proved worthwhile and even necessary to demarcate the target group and their needs, to ensure validity, and avoid biases, the collaboration also proved difficult at times because of their unfamiliarity with and/or skepticism towards standardized tests that are related to the CEFR.

## References

ALTE (forthcoming). *Supplement to the Manual for Language Test Development and Examining: Guidelines for the Development of LSP tests*.

Brunfaut, T. (2014). Language for specific purposes: Current and future Issues. *Language Assessment Quarterly, 11*, 216–225.

Cumming, A. (2013). Assessing integrated skills. In A. Kunnan (Ed.), *The Companion to Language Assessment I* (pp. 216–229). Wiley.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge University Press.

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing, 18*, 171–185.

Gysen, S., & van Avermaet, P. (2005). Issues in functional language performance assessment: The case of the Certificate Dutch as a Foreign Language. *Language Assessment Quarterly, 2*, 51–68.

van der Meulen, M., Hinskens, F., Van der Gucht, F., De Caluwe, J., Heeringa, W., & van der Peet, M. (2016). *Aan het werk met de Staat van het Nederlands! Over de taalkeuze van Vlamingen en Nederlanders in het bedrijfsleven. [Working with Dutch! On language choices in business.]* Retrieved from http://www.meertens.knaw.nl/cms/images/nieuws2016/Rapport_Staat_vh_Ned__Week_vh_Nederlands_Bedrijfsleven_def.pdf

Van Gorp, K., & Deygers, B. (2013). Task-based language assessment. In A. Kunnan (Ed.), *The Companion to Language Assessment II* (pp. 578–593). Wiley.

# Implications of Employing Performance-based Testing in a University Context

**Snezana Mitrovic**, Sapienza University of Rome, Italy

**Abstract:** The issue that this paper addresses is the employment of performance-based testing in a university context, in particular, at the Sapienza University of Rome. The matter is addressed by aiming at designing a performance-based test of the English language. In particular, the issues addressed are: 1) the feasibility of employing a theoretical model of English language knowledge, the Bachman and Palmer one (Bachman & Palmer, 2013) to assess student knowledge; 2) test validation within the university context; 3) the feasibility of employing such a test at the Sapienza University taking into account practical implications (cost-effectiveness, rater training, etc.). To gather data, a questionnaire on personal data as well as two written tasks were designed, accompanied by holistic and analytic rating scales based on the model and the CEFR B2 illustrative descriptors.

## 1 Introduction

There has been an increasing interest in certifying the English language knowledge among Italian high-school and university students. This is mostly due to the fact that Italian universities require a minimum level of English, most often CEFR B2. In order to avoid having to attend a course at the university and pass the university qualifying exam, students often decide to gain a certificate beforehand.

When preparing for an exam, students prepare for the exam format and sample tests are the starting point, which does not necessarily improve their ability to use English in real life. A similar approach is taken when students need to pass the university qualifying exam.

## 2 What is the English language background of an average Italian student?

The elementary school curriculum provides basic education in a number of subjects, including English. English is then taught at all types of upper-secondary school, for five years, from 99 to 132 hours a year, depending on whether it is taught as the first or the second foreign language.

According to *Ministero dell'istruzione, dell'università e della ricerca* (2010a), the Italian Ministry of Education, the following are the aims and objectives of the fifth (last) year foreign language curriculum of lyceums:

> The student acquires linguistic-communicative competences equivalent to the CEFR level B2. The student can produce oral and written texts (in order to report, describe and argue) and reflect on the formal characteristics of texts he/she produces in order to demonstrate an acceptable level of fluency. (p. 16)

The Ministry of Education (2010b, 2010c) sets the same aims and objectives for other types of upper-secondary schools.

## 3 Performance-based assessment of university students

Considering the growing trend towards the certification of the English language, exam specific preparation, as well as the fact that the Ministry of Education sets the objectives but not the means, the question that poses itself is how Italian students would perform on real-life tasks, that is if their performance was assessed and if it would be possible to employ performance-based assessment to evaluate the English language skills of university students.

The questions that the paper addresses are: 1) the feasibility of employing a theoretical model of English language knowledge (Bachman & Palmer, 2013) to design a performance-based test and analytic and holistic scales that would adequately assess the written competence in English of first-year university students of the Sapienza University of Rome; 2) test validation within the university context for the test takers in question; and 3) the feasibility of employing such a test at the Sapienza University or in similar contexts considering the financial and other practical implications such as cost-effectiveness, rater training and availability, etc.

## 4 Theoretical background: testing the ability to use a language

According to McNamara (1996, p. 25), language performance tests developed in response to two main needs: the need to develop selection procedures for foreign students to study at English-medium universities, and 'the need to bring testing into line with developments in language teaching which had resulted from the advent of theories of communicative competence'.

The first performance tests were proposed by Carroll in 1961 and Davies in 1968 in the USA and the UK (McNamara, 1996, p. 24). According to Carroll (1961 [1972, p. 318]) in McNamara 1996, p. 27), language testing is incomplete without integrated performance of examinees. This practically meant that it was essential to determine 'how well the examinee is functioning in the target language, regardless of what his native language happens to be' (Carroll, 1961 [1972, p. 319] as cited in McNamara, 1996, p. 28). It was in this period that the focus moved to the performance on tasks in which different aspects of language knowledge or skills were integrated (McNamara, 1996, p. 28).

## 5 Methodology and sample data

### 5.1 Test and scales design

The methodology employed for gathering information on the learners' English language knowledge is a written criterion-referenced performance test consisting of two parts: writing an enquiry email and a university blog. Each of the test tasks is intended to test the language knowledge at a CEFR B2 level. Analytic as well as holistic scales have been created for each of the tasks. The analytic or multi-trait scales are based on Bachman and Palmer's (2013, p. 45) model of language knowledge and comprise vocabulary, syntax, graphology, cohesion, rhetorical knowledge, functional knowledge, genre and register and knowledge of natural and idiomatic expressions. The holistic scale focuses on the task achievement, that is, completion: to what extent the candidate managed to achieve the task considering all the individual language sub-skills included in the analytic scales. Both scales range from 0 to 4 where 0 equals CEFR A1 level or lower, and 4 equals CEFR B2 level.

The scales have been designed using the CEFR Can Do statements and B2 illustrative descriptors as well as five different course books and online corpora made available by two awarding bodies.

Each test has been rated by two raters, with 10 years of experience in teaching English as a foreign language and working with an awarding body in the area of assessment. The standardization training was done during the pilot sample marking phase.

In addition, a short questionnaire on personal data has been administered, including questions on the age, country of origin, school of origin, study holidays, university qualifying exam, possession of a certificate in English, as well as self-evaluation of English language skills.

### 5.2 Sample data

The test was first administered with a pilot sample, which included 54 second-year university students. Pilot testing confirmed that the tasks elicit the intended sample of language and that the scoring system (scales) is reliable and can be used for consistent marking.

The test was then administered with 186 first-year Sapienza University students, 96.3% Italian students, 96% aged from 18 to 26.

## 6 Test validation

In order to address the issue of inter-rater reliability, the paired sample correlation coefficient for both analytic and holistic scales has been calculated for the pilot sample (the bivariate Pearson correlation coefficient with a two-tailed test of significance for each pair of variables entered): *Task 1 Vocabulary*, *Task 1 Syntax*, *Task 1 Graphology*, *Task 1 Cohesion*, *Task 1 Rhetorical Knowledge*, *Task 1 Functional Knowledge*, *Task 1 Genre and Register*, *Task 1 Natural and Idiomatic Expressions*, *Task 2 Vocabulary*, *Task 2 Syntax*, *Task 2 Graphology*, *Task 2 Cohesion*, *Task 2 Rhetorical Knowledge*, and *Task 2 Natural and Idiomatic Expressions*. The correlation coefficients range from r = .828 to r = .972 (p < .001 in both cases), which indicates a significant positive correlation. The same can be said for the holistic marks: the correlation coefficient r = .943 and r = .939 (p < .001 in both cases) for Task 1 and Task 2 respectively indicate a strong positive correlation.

With regard to the sample, the first-year students, the correlation coefficients for the analytic scale range from r = .861 to r = .962 (p < .001 in both cases), whereas for the holistic scale they are r = .927 and r = .935 (both p < .001) for Task 1 and Task 2 respectively, again indicating a strong positive correlation.

Due to the fact that the administered performance-based test revealed a relatively high variance, Cronbach’s Alpha has been used to estimate the test reliability. The analysis of the pilot sample revealed the reliability coefficient at α = .948 and α = .959 for Task 1 and Task 2 respectively, whereas the sample coefficient at α = .960 and α = .957 for Task 1 and Task 2 respectively demonstrate a high level of internal consistency.

In addition, factor analysis revealed 77% and 75,2% variance for Task 1 and Task 2 respectively explained for the pilot sample, and 73,2% and 74,4% for the first-year university students.

## 7 Results

### 7.1 Student performance

The data collected through the questionnaire have been used to compare the holistic marks for both tasks of different groups of students for each of the independent variables: the age, country of origin, school of origin, whether they have studied abroad, whether they have passed their university qualifying exam and their self-evaluations.

An analysis of variance yielded the following results: the mean values of the students who hold an internationally recognized certificate in English ($\bar{x}$ = 2.36 and $\bar{x}$ = 2.33 for Task 1 and Task 2 respectively) is greater than the mean values of the ones who do not ($\bar{x}$ = 1.66 and $\bar{x}$ = 1.84 for Task 1 and Task 2 respectively). In the same way, it is greater for the students who have studied abroad ($\bar{x}$ = 2.16 and $\bar{x}$ = 2.24 for Task 1 and Task 2 respectively against $\bar{x}$ = 1.75 and $\bar{x}$ = 1.85 who have not) as well as for the ones who have passed the university qualifying exam in English ($\bar{x}$ = 2.13 and $\bar{x}$ = 2.32 for Task 1 and Task 2 respectively against $\bar{x}$ = 1.78 and $\bar{x}$ = 1.81 who have not).

Furthermore, Kendall’s Tau-b correlation coefficient of $\tau$ = .419 indicates a moderate positive relationship between the students’ self-evaluation of English language knowledge and their average holistic mark on the writing test.

The mentioned independent variables are the ones that positively influence the dependent ones. The rest of the data collected through the questionnaire did not prove significant for the student performance.

### 7.2 CEFR B2: an attainable goal?

Converted into CEFR levels (where 1 is CEFR A1 or lower and 4 CEFR B2) and based on the average holistic mark across the two tasks, the students’ marks mostly fall under CEFR A2, 37%, while the level of English of 31% of the students in the sample demonstrated a CEFR B1 level, 23% CEFR B2 level and 9% A1 or lower.

## 8 Performance-based assessment: yes or no?

### 8.1 Implications: difficulties of the approach

#### *8.1.1 Model applicability*

Since the model used for the scale design and assessment is the Bachman and Palmer (2013) one, its analytical nature implies the assessment of each individual component of the model. The most obvious disadvantage of the model is that not each of the model components can be evaluated by a single task. With regard to the tasks administered, an enquiry email and a blog entry, designing appropriate descriptors for some of the components has proven to be a challenge. For example, online communication (an enquiry email) does not necessarily require a high level of formality. In the same way, there is no fixed format for blog entries, which again made the evaluation of genre and register difficult.

#### *8.1.2 Student availability*

The test was administered in one of the university faculties, during regular lessons and was not mandatory for the students. For this reason, as well as due to the time constraints, it was impossible to administer the test with a larger sample.

The University benefit of the administration of this particular test was to examine the English language level of their students and measure it against their university qualifying exam. If this approach was to become standard practice, student availability would not pose itself as a problem.

#### *8.1.3 Cost-effectiveness and raters*

Unlike standardized language tests, with multiple-choice questions, where marking is done automatically, performance-based assessment requires an analytic evaluation of skills. This of course requires more time and assets such as trained raters, who need to go through a standardization process. If this approach was to be replicated in a university context as common practice or to replace the university qualifying exam, it would imply costs that would be much higher than the ones of standardized language tests due to the fact that it is time-consuming and that it requires trained and experienced raters, and raises issues such as inter-rater reliability.

### 8.2 Implications

#### *8.2.1 Student strengths and weaknesses*

The most evident beneficial traits of the approach are that it has potential washback effect in small scale assessment, such as university context as well as that it allows for the identification of student strengths and weaknesses.

Extended-production responses provide valuable detailed information about student knowledge. In the case of first-year university students of the Sapienza University, a significant weakness is the negative transfer from Italian and consequently appropriacy or cultural issues. For example, questions such as ‘Is there a college where I can sleep with other students?’

The differences between the holistic marks for the two tasks indicate that one of the students’ strengths is their ability to rely on the input and use the information they were provided with. The input for Task 1 was considerably longer and provided language for the students to rely on, which the students used in their responses. Consequently, the level of achievement is higher. Another reason for this is the fact that despite their obvious limitations and independent of the mastery of the sub-skills, the students did manage to communicate the message.

The holistic marks awarded for Task 2 are considerably lower. This is mostly due to the fact that very few students actually wrote a well-organized and convincing article with original ideas and a specific point of view. The input contained very short, specific instructions and for that reason the students needed to provide content themselves.

### *8.2.2 Analytic scales – analytic marking and holistic positive assessment*

Finally, the analytic approach based on analytic scales grants assessment of each of the model components or language sub-skills. A positive holistic approach to marking, based on Can Do statements on the other hand, evaluates student knowledge based on what they can do not what they cannot do and prioritizes their strengths over their weaknesses. The use of the two types of scales together provides more information about the student knowledge than the use of a single scale or standardized tests.

## 9 Conclusion

With regard to the use of a theoretical model of language knowledge, it is evident that it is not universally applicable and that it requires certain modifications depending on the task to which it needs to be applied and the context in which it is employed.

Despite the approach disadvantages, such as cost-effectiveness and difficulties with the design of some of the descriptors for the analytic scales, the advantages of this kind of approach, especially the potential washback effect, are quite significant. Whether the advantages outweigh the disadvantages would depend on a number of factors; however, in small-scale assessment this kind of approach is certainly feasible and beneficial.

## Further Reading

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1982). The Construct Validation of Some Components of Communicative Proficiency. TESOL Quarterly, 16(4), 449-65. doi:10.2307/3586464

Brown, J. D. (2002). The Cronbach alpha reliability estimate. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *6*(1): 17–18.

Council of Europe. (2001a). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. Retrieved from http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

Council of Europe. (2001b). *Manual for Language Test Development and Examining: for use with the CEFR*. Strasbourg: Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A Manual*. Strasbourg: Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf

Fulcher, G. (2010). *Practical language testing*. London, UK: Hodder Education.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217–262. doi:10.1177/026553229801500204

North, B. (2007). The CEFR Illustrative Descriptor Scales. *The Modern Language Journal, 91*(4), 656–659. doi:10.1111/j.1540-4781.2007.00627_3.x

Weigle, S. (2002). *Assessing Writing*. Cambridge, UK: Cambridge University Press.

Weir, C. J. (2005). Language Testing and Validation: An Evidence-Based Approach. Basingstoke, UK: Palgrave Macmillan.

## References

Bachman, L. F., & Palmer, A. S. (2013). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.

McNamara, T. F. (1996). *Measuring Second Language Performance*. London, UK: Longman.

Ministero dell'istruzione, dell'università e della ricerca. (2010a). *Indicazioni nazionali riguardanti gli obiettivi specifici di apprendimento concernenti le attività e gli insegnamenti compresi nei piani degli studi previsti per i percorsi liceali*. Retrieved from http://www.indire.it/lucabas/lkmw_file/licei2010/indicazioni_nuovo_impaginato/_decreto_indicazioni_nazionali.pdf

Ministero dell'istruzione, dell'università e della ricerca. (2010b). *Il regolamento degli istituti professionali*. Retrieved from http://archivio.pubblica.istruzione.it/riforma_superiori/nuovesuperiori/doc/Regolam_professionali_04_02_2010.pdf

Ministero dell'istruzione, dell'università e della ricerca. (2010c). *Il regolamento degli istituti tecnici*. Retrieved from http://archivio.pubblica.istruzione.it/riforma_superiori/nuovesuperiori/doc/Regolam_tecnici_def_04_02_10.pdf

# Testing Pre-service Teachers' Spoken English Proficiency

**Odette Vassallo**, University of Malta, Malta
**Daniel Xerri**, University of Malta, Malta
**Sarah Grech**, University of Malta, Malta

**Abstract:** This paper discusses the recent introduction of the Spoken English Proficiency Test for Teachers (SEPTT) in Malta. By means of this test, the regulator of the English Language Teaching (ELT) industry in the country is seeking to ensure high levels of spoken English proficiency amongst pre-service teachers. SEPTT tests candidates' ability to use spoken English for a variety of functions, including conversing, explaining, presenting information, giving instructions and responding in a context specific to teaching. The paper shows how the test is based on the idea that people require different proficiencies that are always situated in particular contexts and bounded by a particular social practice (Freeman, Katz, Garcia Gomez, & Burns, 2015). Hence, an ESP-derived approach to language proficiency is required whereby teacher education focuses on the specific linguistic needs of teachers when enacting their role. This paper describes the research that went into SEPTT's design and implementation.

## 1 Introduction

The term 'proficiency' is hard to define, especially given its various uses in assessment. However, there appears to be a shared understanding of the qualities a proficient user should possess. These are competence and skill in the target language. As a result of the fast changing realities of the ELT industry and an increased number of English language speakers globally, a growing concern about teachers' spoken proficiency has become more pronounced than ever before. Varieties of English are bound to generate diverse perceptions when evaluating oral communication or ensuring an adequate classroom model for learners. Thus, a teacher's spoken production in the target language has become central to the ELT industry in Malta. Operational as from 2017, SEPTT is designed to ensure high levels of spoken English proficiency amongst ELT practitioners, specifically, pre-service teachers working in a variety of international contexts. A determining factor in the design of SEPTT was the decision to move away from 'native-speakerism' (Holliday, 2006). In fact, SEPTT does not adopt a normative standard based on the notion of native speaker.

Apart from incorporating the standard assessment criteria typical of the speaking component in a general English proficiency test, SEPTT introduces a new dimension; more precisely, it tests the language teachers are expected to use in a classroom context. Based on Freeman et al.'s (2015) English-for-Teaching, the test materials are modelled on teachers' use of English in classroom discourse. This is represented by a fifth criterion in the test's rating scale, which is teacher discourse. SEPTT tests candidates' ability to use spoken English for a variety of functions, including conversing, explaining, presenting information, and giving instructions and feedback in a context specific to ELT.

## 2 English for teaching

SEPTT is based on the notion that people require different proficiencies that are always situated in particular contexts and bounded by a particular social practice (Freeman, 2015). Just as general English proficiency cannot address all the linguistic needs of students in the world beyond the classroom, it cannot fully specify the demands on teachers' use of language inside the classroom when teaching the language. Equally important is the attention that should be given to language use in teacher talk and classroom discourse. Pre-service teachers are thus made aware that by enhancing language use they could increase students' learning potential (Walsh, 2002).

Walsh's (2002) study focuses on the "relationship between language use and pedagogic purpose" (p. 4); he stresses that control of the use of language is as important as the choice of methodologies. He contends that "the teacher by controlled use of language and by matching pedagogic and linguistic goals, facilitates and promotes reformulation, clarification, leading to greater involvement and precision of language on the part of the learners" (Walsh, 2002, p. 4). Van Canh and Renandya (2017) echo the importance attributed to language use by emphasising the need for the ELT practitioner to not only be highly proficient in general English but also "adept at using the language to create conducive learning environments" (p. 79).

Having factored in all the reasons for encouraging a more intense focus on teacher discourse, an ESP-derived approach to language proficiency is required whereby teacher education focuses on the specific linguistic needs of teachers when enacting their role. According to Freeman et al. (2015), such a "focused approach converts the problem of language improvement from one of general proficiency to one of specialized contextual language use, which is likely to be more efficient in bringing about practical impacts on teacher classroom efficacy and student learning outcomes" (p. 131). This acts as the foundation for the construct of English-for-Teaching, i.e. the essential language skills needed to prepare and enact a lesson in English (Young, Freeman, Hauck, Garcia Gomez, & Papageorgiou, 2014). One of the main implications of English-for-Teaching is that teacher language assessment needs to change so that the focus is on classroom-specific language proficiency rather than general language proficiency. This is imperative given that "Creating assessments that actually look like the work teachers do in the classroom can help build stronger validity arguments" (Freeman et al., 2015, p. 138).

A similar approach to assessment was developed by Douglas (2001) who asserts that Language for Specific Purposes (LSP) tests "derive their content from an analysis of specific language use situations of importance to the test-takers" (p. 172). SEPTT was designed after analysing the target language use (TLU) (Douglas, 2001); teacher discourse is the criterion that takes into account language use in a specific purpose context. The target situation is specific to pre-service teachers whose classroom experience is limited to the teaching practice sessions held during their pre-service training course.

Some of the tasks in SEPTT replicate the classroom tasks and routines that teachers typically engage in. Freeman et al. (2015) inventoried these tasks and routines and grouped them into three functional areas: managing the classroom; understanding and communicating lesson content; and assessing students and giving them feedback. The proficiency construct in SEPTT is framed by the context in which teachers typically use English when teaching the language. Hence, SEPTT is only appropriate for candidates who have completed a pre-service ELT methodology course. Despite the fact that candidates' knowledge of methodology is the basis on which their spoken English proficiency is tested, the knowledge itself is not assessed in SEPTT.

## 3 Context

English has official status in Malta. It is given significant importance in various domains and it is studied throughout compulsory education. The target test takers are either first or second language speakers of English. Malta is one of a handful of countries to have legislation in place to regulate teaching in the ELT industry, which is responsible for over 75,000 international students per year. As the regulatory body, the ELT Council strives to maintain high standards in ELT qualifications and is responsible for issuing teaching permits. It operates quality assurance systems in all aspects of the ELT industry; this includes periodic monitoring visits to ensure compliance with established quality standards.

SEPTT is a legal requirement for teachers applying for a teaching permit in Malta. In addition to SEPTT, prospective teachers applying for a permit should be in possession of an Advanced level certificate in English or a language awareness qualification, together with a pre-service teacher training qualification. Although the Advanced level certificate in English and language proficiency test incorporate a speaking component, they were never designed to assess teacher discourse. SEPTT is the final examination pre-service teachers sit prior to obtaining a teaching permit; therefore, candidates would have already obtained qualifications in general English proficiency and teacher training. Such a combination of qualifications is fundamental to the selection and development of the test material as it is grounded in candidates' knowledge of pedagogy.

## 4 SEPTT design

Stakeholder representatives were part of the task design team and these ensured that the test is both comprehensible to teachers, and acceptable to stakeholders (see Andrews, 2004). The introduction of SEPTT underwent a process of public consultation as part of a revised legal notice. Following this, the test was designed and a detailed manual produced. Upon completion of all the documentation, a more focused consultation exercise was conducted with school owners, who bear most of the impact, and schools' Directors of Studies, who experience the test's washback effect through teacher training and teacher recruitment. Finally, intensive training sessions were organised for the examiners.

### 4.1 The three-part structure

The test content is authentically representative of tasks conducted by teachers in the target situation. SEPTT is divided into three tasks and takes no longer than 15 minutes. The first part serves the purpose of establishing the role of the candidate as a prospective teacher. This is followed by a gradual increase in the challenge of the second and third task, the long turn and conversation respectively. Both parts aim to immerse the candidate in a teaching-related situation and their language use is tested. It is important to establish that SEPTT does not test knowledge of pedagogy but instead it exploits that knowledge to elicit teacher discourse based on the activities determined by the test materials.

### *4.1.1 Part 1 – Interview: the teacher*

Part 1 consists of an introductory interview in which the examiner asks questions about the candidate's interests, plans, and training in relation to ELT. Questions may also focus on the candidate's views about teachers, teaching and learning.

The task takes the form of a two-way exchange initiated by the examiner where the candidate is expected to respond to a set of questions. The questions posed by the examiner may focus on past, present or future situations.

The task is aimed at assessing candidates' ability to provide information about familiar topics related to ELT, as well as details about their interests with respect to this profession.

### *4.1.2 Part 2 – Long turn: the lesson*

This part is a three-minute presentation expressed as a long turn by the candidate based on a prompt focusing on some aspect of an English language lesson, such as managing the classroom, communicating content, or setting up an activity. Before entering the test room, the candidate is provided with 10 minutes in which to examine the prompt. Prior to the presentation, the candidate is provided with three minutes for preparation.

This task focuses on extended, structured speaking using a prompt that clearly outlines what the candidate is meant to do in a particular scenario. Besides a detailed rubric, the prompt might also include printed or visual components that would aid the candidate in the delivery of the presentation.

This task assesses the candidate's fluency and accuracy in presenting, defining, developing and exploring information related to the prompt. The presentation needs to include a description and explanation of what the candidate would do and why.

When delivering the presentation, the candidate might need to: introduce the presentation by indicating how each part of the prompt will be discussed; define and focus on each part of the prompt; exemplify each part of the prompt; conclude the presentation by summarizing, referring to future situations, identifying main areas of concern, suggesting the course of action required, or indicating personal experiences and views.

### *4.1.3 Part 3 – Conversation: instructions and response*

In the third and final part, a conversation between the examiner and candidate takes place, which is based on a given scenario related to the prompt in Part 2. The candidate is provided with a rubric and one minute in which to examine it. Then the candidate is asked a number of questions.

This task focuses on the candidate's ability to respond to a particular lesson scenario or provide instructions to learners. The candidate might be asked to describe how s/he would address a specific situation or what kind of instructions they would provide to learners. The candidate is expected to use concrete examples when answering. By means of a set of questions, the candidate may be required to speculate, evaluate, compare and contrast, explore possibilities, extend situations and experiences, and suggest alternative perspectives.

### 4.2 Measurement of performance

A candidate's performance on SEPTT is assessed by means of an analytic rating scale made up of five criteria and 20 descriptors corresponding to four bands, Band 4 being the highest level of proficiency and Band 1 being the lowest. The five criteria are: teacher discourse; coherence and cohesion; pronunciation; grammar; and vocabulary. At the end of the test, the examiner determines the band that best describes a candidate's performance with respect to each criterion across all three tasks. The lowest band attained for a specific criterion determines the global band attained in the test. A global Band 4 and 3 allows the candidate to obtain a teaching permit.

The examiner in SEPTT also acts as an interlocutor and is responsible for timing every single part of the test, initiating interaction with the candidate, and assessing the candidate's performance. Interaction with the candidate, including the instructions provided to the candidate and all the questions posed to the candidate, is scripted for the purpose of ensuring reliability. Every SEPTT examiner is periodically provided with rigorous training on how to follow test procedures in a consistent manner, and on how to interpret the rating scale for the purpose of reliably assessing candidates' performance. Every single test is recorded and these recordings are used to regularly monitor examiners' rating performance.

## 5 Conclusion

As was to be expected, the implementation of SEPTT generated some level of anxiety among teachers and trainers; this was less evident with Directors of Studies. One group resisted the test, another one ignored it until it was launched, and a third group embraced it as a welcome change (for similar responses in other studies, see Andrews, 1994).

The first data set was collected and went through a preliminary analysis. The test tasks, instructions, materials and scoring method seem to be producing the desired results. SEPTT allows for the possibility 'to make inferences about a test taker's capacity to use language in the specific purpose domain' (Douglas, 2000, p. 19). Different sessions were compared to one another and moderation was conducted following each session.

Research is a key factor in evaluating both the intended and unintended consequences of SEPTT; thus, further collaboration and consultation with stakeholders is currently underway. The research that is being conducted on SEPTT consists of an investigation into the impact and washback of SEPTT on the ELT industry at large, on school management, on teacher trainers, on pre-service teachers, and on teaching and learning. Some of the research projects that are presently being carried out involve:

- interviews with teacher trainers from different pre-service training courses;
- interviews with Directors of Studies regarding new recruits' spoken proficiency following SEPTT's implementation;
- observation of teacher trainers during pre-service training courses while focusing on developing trainees' spoken proficiency;
- investigating whether an impact on teachers' methodology is taking place;
- investigating whether teachers have changed their instructional practices since SEPTT's implementation, and if so, whether this has effected a change in student learning.

The long-term plan is to analyse the speech data collected from each examination session with a view to comparing the teacher discourse used in SEPTT with authentic classroom discourse.

Since becoming operational in 2017, it is already evident that SEPTT is having an effect on pre-service teachers, trainers and school management. This is because the test is a legal requirement in Malta and has a bearing on teacher training and recruitment. SEPTT is a high-stakes test that determines whether a pre-service teacher obtains a teaching permit or not. Teacher trainers who are responsible for pre-service courses are now bound to dedicate classroom time to prepare trainee teachers for SEPTT. By shifting the spotlight onto oral communication in the classroom and building a strong association with language use in teacher talk as an essential part of classroom discourse, it is possible that SEPTT is elevating teachers' spoken English proficiency to the same level of importance as the ELT methodologies taught in teacher training courses.

## References

Andrews, S. (1994). Washback or washout? The relationship between examination reform and curriculum innovation. In D. Nunan, R. Berry & V. Berry (Eds.), *Bringing about change in language education* (pp. 67–81). Hong Kong: Department of Curriculum Studies, University of Hong Kong.

Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 37–50). Mahwah, NJ: Erlbaum.

Douglas, D. (2000). *Assessing language for specific purposes: Theory and practice*. Cambridge, UK: Cambridge University Press.

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing, 18*(2), 171–185.

Freeman, D. (2015, April). *Frozen in thought? How we think about what we do in ELT*. Keynote address at the 49th Annual International IATEFL Conference and Exhibition, Manchester, UK.

Freeman, D., Katz, A., Garcia Gomez, P., & Burns, A. (2015). English-for-Teaching: Rethinking teacher proficiency in the classroom. *ELT Journal, 69*(2), 129–139.

Holliday, A. (2006). Nativespeakerism. *ELT Journal, 60*(4), 385–87.

Van Canh, L., & Renandya, W. A. (2017). Teachers' English proficiency and classroom language use: A conversation analysis study. *RELC Journal, 48*(1), 67–81.

Walsh, S. (2002). Construction or obstruction: Teacher talk and learner involvement in the EFL classroom. *Language Teaching Research, 6*(1), 3–23.

Young, J. W., Freeman, D., Hauck, M., Garcia Gomez, P., & Papageorgiou, S. (2014). *A design framework for the ELTeach program assessments*. Princeton, NJ: Educational Testing Service.

# Validating University Entrance Policy Assumptions. Some Inconvenient Facts.

**Bart Deygers**, KU Leuven, Centre for Language and Education, Belgium

**Abstract:** It is common practice at universities worldwide to use minimal language requirements to discriminate between L2 students who are allowed to register, and those who are not. In Flanders, Belgium, international L2 students are required to demonstrate B2 proficiency in order to be eligible for university admission. There are a number of commonly accepted ways in which these students can prove B2 Dutch language proficiency, but the use of these requirements is based on a number of unverified assumptions. It is the purpose of this paper to identify and examine these assumptions, and to present the combined findings of a doctoral research project. Readers who are looking for more detail will find references to more detailed publications at the end of this paper.

## 1 University admission language requirements: common, and commonly unquestioned

It is common practice for institutions of higher education to set minimal language requirements for aspiring L2 students and use language tests as proof of that level. Typically, the use of such language requirements to control university admission relies on the assumption that successful linguistic participation in academia demands a certain language level. In Europe this level is typically set at B2 of the Common European Framework of Reference (CEFR, Council of Europe, 2001) (Deygers, Zeidler, Vilcu, & Carlsen, 2017). For most L2 students, registration is conditional on proving B2 proficiency, while the L1 population is exempted from any language requirements.

Strikingly, the assumptions which support a university admission policy have not been the subject of much empirical research (McNamara & Ryan, 2011). In most contexts, the use of language requirements for university admission is accepted as common practice, without questioning the premise on which this practice relies. The doctoral research on which this contribution reports did question a number of these key policy assumptions.

## 2 University admission requirements in Flanders, Belgium

The goal of the Flemish university admission policy is to select students who have a sufficient level of Dutch language proficiency to be able to attend a Dutch-medium university program. The minimal level at which L2 students can be presumed to successfully partake in the various linguistic challenges at university was set at B2. In line with the CEFR, a B2 learner can be described as somebody who can understand the main ideas of complex texts, interact fluently and spontaneously with native speakers, produce clear and detailed texts, and develop a sustained argumentation (Council of Europe, 2001, p. 24).

The university admission policies of the five Flemish universities all list three primary ways in which applicants can prove B2 ability. The most commonly used way is to pass one of two B2 tests. Alternatively, they can present the admission officer with a certificate that shows successful completion of one year at a Dutch-medium secondary school. Thirdly, they can register for a Dutch-medium university program if they have successfully completed one year in Dutch-medium higher education, which implies that these candidates had already proven B2 proficiency the year before.

Supporting these requirements, are four assumptions which are implicitly or explicitly present in the policy texts. These assumptions have also been checked with and confirmed by all relevant policy makers. In what follows, each assumption will be introduced, and a summary of the research findings will be given.

## 3 Examining assumptions

### 3.1 Assumption 1: B2 constitutes an adequate threshold level to determine international L2 students' access to a Dutch-medium university in Flanders

As is the case in most European countries, the default proficiency level used in the context of international L2 student university admission is B2, but the rationale for its widespread use is rather thin or even non-existent (Deygers et al., 2017). In a research project that investigated whether L2 applicants with B2 proficiency can be considered ready for the linguistic demands of university, academic staff, and L2 students were consulted.

When confronted with language performance samples, university staff (N = 24) considered the B2 level vastly insufficient for listening and reading, but acceptable for writing. Speaking skills were considered non-essential for first-year students. Similarly, international L2 students (N = 20) who were tracked during their first year at university all struggled with the actual listening demands of university. All L2 students reported problems understanding their first lectures, mainly because they were not prepared for the variation in accents and pronunciation styles encountered in real-world lectures. The international L2 students reported fewer problems with reading and writing, primarily because these skills typically allow language learners to deal with input at their own pace. Reading in Dutch was estimated to take twice as long as compared to reading in the L1. Additionally, L2 students who performed well on a language test did not necessarily do well academically. In line with previous research (e.g., Cho & Bridgeman, 2012; Ingram & Bayliss, 2007), quantitative analysis showed a weak, non-significant relationship between the scores on both accepted language tests and academic success (Test 1: W = 46, p = .625, effect size r = -.115; Test 2: W = 51, p = .599, effect size r = -.120).

This study, in short, found no evidence to support the assumption that a uniform B2 requirement provides an adequate threshold level to make claims about students' ability to meet the linguistic requirements of academia. Perhaps, as Hulstijn (2011) suggests, there is more merit in using differentiated CEFR-based requirements that are in line with the actual real-world needs.

### 3.2 Assumption 2: the B2 tests are equivalent measures of B2 proficiency

At every university, two B2 tests are considered equivalent measures of B2 proficiency, but without a clear rationale or empirical support for this presumed equivalence. The research conducted in this doctoral research project shows that the fact that both tests have been properly linked to the same CEFR level does not guarantee equivalence.

The correlation between the overall scores and the writing scores ($N = 118$) on both tests was moderately high (overall $r = .767$, $p < .001$; writing $r = .694$, $p < .001$), but the agreement between the scores on the oral tests was much lower ($\tau = .387$, $p < .001$). T-tests confirmed that the differences between mean scores were significant ($p < .001$), with effect sizes ranging from $d = -0.53$ (writing components) to $d = -1.41$ (speaking component). Additionally, the pass probability was found to be significantly ($p < .05$) different (50% vs. 35%). A closer study of the item-level scores showed not only that the scores differed, but also why they did. The two tests stressed different components in a different way, leading to a very different conceptualization and weighting of grammar and vocabulary. Strikingly, even when these tests used the same CEFR-based criteria to assess the same task types, no agreement was to be found.

All in all, there was little or no evidence to support the presumption of equivalence. The two tests will quite likely judge candidates with a distinctly high or low proficiency quite similarly, but may differ substantially in their assessment of candidates with a less clear profile.

### 3.3 Assumption 3: Flemish high school graduates have B2 proficiency

To examine whether all students with a Flemish high school degree have attained the B2 level in Dutch, 159 first-year Flemish L1 students sat two written STRT tasks during their first month of university education. Using non-parametric statistics and Multi-Faceted Rasch analysis, the L1 scores were compared against the performance of two groups of L2 candidates: L2 students who had studied Dutch at their home institution ($N = 629$), and L2 students who had done so in Flanders ($N = 116$). The results showed that L1 students significantly ($p < .000$) outperformed both groups of L2 students – both overall and on the linguistic criteria of both tasks (when using a conglomerate score for all linguistic criteria used in one task), with medium effect sizes. L2 students who had studied Dutch abroad achieved significantly ($p < .000$) higher scores on content criteria. The L2 students who had studied Dutch in Flanders were the lowest-scoring group.

Importantly, however, 11 percent of the L1 students did not attain the B2 level as measured by the writing tasks on the L2 test. Logistic regression showed that out of all linguistic criteria, scores on Grammar and Vocabulary were the best predictors of membership to the group of Flemish students. Consequently, assuming that Flemish high school graduates will have B2 proficiency in Dutch is mostly true, but should not be considered self-evident.

### 3.4 Assumption 4: L2 gains will be made during the first year at university

In order to measure the language gains and document the experiences of international L2 students at Flemish universities, 20 respondents were regularly interviewed during their first academic year at a Flemish university. After eight months, the respondents who had not left university ($n = 13$) took two writing and speaking tasks again (the combination of these tasks was predictive for the overall score at R_adj^2 = .908, $p < .000$).

The results showed that the respondents had made no significant gains in terms of test score, or in terms of measures of complexity, accuracy, or fluency. The interview data showed that nearly all respondents had experienced some degree of social and academic isolation, and reported a perceived lack of institutional support. Likely, an important reason why the respondents had made zero gains was the limited meaningful interaction with L1 speakers.

### 4 Conclusion

The results of this research project partially or fully disprove the main claims that support the Flemish university entrance policy, and shed doubt on its effectiveness. First, at B2, the minimum performance level is below the real-life listening requirements. Furthermore, the university admission policy is unlikely to guarantee a consistent minimum language level among students, since the tests used cannot be considered equivalent, and since people who are exempt from taking the test are not sure to pass it. Lastly, international L2 students who do enroll after passing the entrance test make very few language gains during their first year, and do not easily gain access to the academic community.

There are a number of possible ways in which this dissertation could have real-world impact, but perhaps the most important recommendation concerns not the entrance requirements, but the post-admittance policy. When international L2 students register for Dutch-medium programs, their language proficiency is not quite high enough to meet the real-world demands, yet no specific accommodations are in place for this group. To recognize the presence of this group, to create the circumstances that would help them build a network, and to address their language-related needs, would be a big step forward. Universities could also help international L2 students make language gains by providing curricular language classes that offer language support throughout their academic trajectory.

### Further reading

For a more detailed discussion of the research data and findings, please consider reading the following publications:

Deygers, B. (2017, accepted). A year of highs and lows. Considering contextual factors to explain L2 gains at university. *The Modern Language Journal*.

Deygers, B., Van den Branden, K., & Peters, E. (2017). Checking assumed proficiency: Comparing L1 and L2 performance on a university entrance test. *Assessing Writing, 32*, 43–56.

Deygers, B., Van den Branden, K., & Van Gorp, K. (2017, published online). University entrance language tests: a matter of justice. *Language Testing*.

Deygers, B., Van Gorp, K., & Demeester, T. (2017, in press). The B2 level and the dream of a common standard. *Language Assessment Quarterly*.

Deygers, B. (2017, in press). University entrance language tests: examining assumed equivalence. In J. Davis, J. Norris, M. Malone, T. McKay, & Y. Son (Eds.). *Useful Assessment And Evaluation In Language Education*. Washington, DC: Georgetown University Press.

## References

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing, 29*(3), 421–442.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg: Council of Europe.

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2017). One Framework to Unite Them All? Use of the CEFR in European University Entrance Policies. *Language Assessment Quarterly*. 1–13.

Hulstijn, J. H. (2011). Language Proficiency in Native and Nonnative Speakers: An Agenda for Research and Suggestions for Second-Language Assessment. *Language Assessment Quarterly, 8*(3), 229–249.

Ingram, D., & Bayliss, A. (2007). *IELTS as a predictor of academic language performance*, Part 1 (Vol. 7). Melbourne.

McNamara, T., & Ryan, K. (2011). Fairness Versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8*(2), 161–178.

# Testing the Test: How Politics Influenced the Reception of an English Test for Lecturers

**Frank van Splunder**, University of Antwerp, Belgium
**Catherine Verguts**, Ghent University, Belgium

**Abstract:** The paper discusses the implementation and perception of the Interuniversity Test of Academic English (ITACE) for lecturers, a test which was developed in response to the Flemish Government's decision to test the C1-level of lecturers teaching in English in Flemish higher education. The case study reveals how the implementation of the ITACE determined its reception. Implemented top-down, the test sparked a media storm in which the test became the scapegoat. Its very purpose – quality assurance in higher education – was largely neglected in the media. In the paper we will discuss the press coverage, and we will argue why it was inaccurate. We will show how the implementation of the test and the press coverage were counterproductive to the acceptance of the test as a means to ensure the quality of teaching and improve employability. The case study reveals that language testing is more than testing language.

## 1 Introduction

The overarching context of this research is globalization and, as a consequence, the increasing use of English as a lingua franca. As may be observed in European higher education, more and more courses are taught in English (Wächter, 2014). This trend can be traced back to the 1999 Bologna Declaration and the resulting European Higher Education Area. Even though 'Bologna' stressed the importance of different languages and cultures, English emerged as the preferred language in higher education, most notably its use as a medium of instruction. This trend appears to be informed by the forces of globalization and neoliberalism, which prefer the use of a single language (Block, Gray & Holborrow, 2012). Moreover, the marketing of English as the preferred medium of instruction fits in with the commodification of education. It may be no surprise that English language testing has become big business too. Most of these tests have been developed in English-speaking countries, particularly in the United States (e.g. TOEFL) and in the United Kingdom (e.g. IELTS). These tests are clear cases of high-stakes tests with social as well as commercial implications (Spolsky, 2012, p. 497–499).

In today's higher education in Europe, one may observe a trend to promote one's national language (e.g. Dutch), while at the same time introducing an international language (which in practice means English) in an increasingly multilingual and multicultural context (often referred to as 'superdiversity') (van Splunder, 2016, p. 209). The introduction of English as a lingua franca in higher education challenges the 'ownership' of English (Widdowson, 1994). It addresses questions such as: Who makes the rules (grammars, textbooks, etc.)? Who makes the language tests? Even though this lucrative business may still be the realm of the native speaker, non-native speakers are increasingly claiming their share of the cake, reflecting the idea of English as a lingua franca which belongs to an international community rather than to a much smaller community of native speakers (Jenkins, 2007; Seidlhofer, 2011).

Apart from a testing purpose, language tests may have a considerable impact on an individual or on a society at large. To put it rather bluntly, language testing is more than testing language. As tests select and inherently discriminate, passing a high-stakes test may yield important benefits, while failing a test can jeopardize one's future. Tests may be used to achieve political goals, resulting in “the politicization of assessment” (McNamara & Roever, 2006, p. 213). As a result, tests reflect the values and beliefs held by its stakeholders, including, for instance, politicians (Fulcher, 2009, p. 5–8). As observed by Shohamy (2006), tests can be employed as a means to create and maintain social order. They may be used to control educational systems and even societies in that they define what kind of knowledge is required (Shohamy, 2001) or what standards should be set (e.g. the CEFR). On the other hand, tests can also serve as a means to empower people, as they are a means to gain access to education or employment. Spolsky (2012, p. 503) points out that language tests can serve as “an excellent instrument for intelligent and responsible language management”, whereas its misuse reflects flawed language policy. Thus, whereas a language test can be used as a gatekeeper, it can serve as a gate opener as well (Bachman & Purpura, 2008, p. 456).

## 2 Language policy and testing in Flanders

The current paper discusses the implementation and perception of the Interuniversity Test of Academic English (ITACE) for lecturers, a test which was developed in Flanders, the Dutch-speaking part of Belgium. Flanders has a large degree of autonomy within Belgium, for instance in educational matters, as a result of which it has developed its own language policy. In 2013, the Flemish Government decided that all lecturers in higher education needed to prove their language skills if they wanted to teach in a language different from their mother tongue, which in practice meant that all lecturers teaching in English curricula needed to prove their C1-level of the CEFR. C1 is the advanced level needed for social, academic and professional purposes (Council of Europe, 2001).

Flanders may be the only region in Europe which has developed its own English language test for lecturers in higher education. The Government’s decision to implement a mandatory language test reflects the ideology that language should be regulated from above. An ideology may be described as a ‘set of beliefs’ which may be either manifest or latent (Barakos, 2016, p. 38). The belief that language should be regulated is firmly held in Belgium, where almost everything is divided along linguistic lines, and language laws are a major issue. Whereas in the past, language laws served to negotiate relations between the two major languages in Belgium (Dutch and French), in an increasingly international academic context English emerged as the ‘other’ language. This may be observed particularly in Flanders, where a law was passed concerning the use of languages in higher education (Flemish Decree Concerning Language Regulation in Higher Education, 2012; for a more detailed account: see van Splunder, 2016). The implementation of a mandatory language test reflects Flemish language sensitivity as well as a strong tradition of top-down language control. Language issues are also widely debated in the media.

The implementation of a language test revealed conflicting discourses between various stakeholders. As stated earlier, the test results from the Flemish Government’s language policy, which states that anyone willing to teach in a language other than their native language should take a language test. In most cases, this ‘other’ language happens to be English. Apart from internationally recognized tests such as IELTS or TOEFL, the Government also accepted the

ITACE, which was designed as an academic language test for lecturers. Although initially set up to comply with legal requirements, the Interuniversity Testing Consortium (IUTC), which designed the ITACE, made possible a unique collaboration between several Flemish universities in bringing together their expertise in language testing. Currently, four Flemish universities are collaborating in the project: Ghent University (Universiteit Gent), the University of Leuven (KULeuven), the University of Antwerp (Universiteit Antwerpen), and the Vrije Universiteit Brussel (VUB). The test comes in two versions: one for lecturers and one for students (http://www.itace.be). The focus of this research is the ITACE for Lecturers, which consists of three parts: an online test (reading, listening, grammar, vocabulary), a writing test (based on the test taker's field of research), and an oral test (based on the test taker's teaching practices).

Apart from devising a reliable and standardized language test which would meet international criteria, the aim was to devise a test with considerable advantages when compared to other tests: the ITACE is purpose-built as it serves a particular audience (lecturers teaching in English as their additional medium of instruction), it is domain-specific, and it is flexible in use. Moreover, the Consortium remains in control and it cashes in as well. The overarching objectives are to improve the lecturers' level of English and to ensure the quality of education. Following its introduction, however, the mandatory language test sparked a media storm in which the ITACE became the scapegoat. The test was largely seen as a political tool of the Government, and its very purpose was largely neglected in the media. In this paper we will discuss the press coverage, and we will argue why it was inaccurate. We will show how the top-down implementation of the test was counterproductive to the acceptance of the test as a means to ensure the quality of teaching and improve employability.

## 3 Case study: the ITACE in the media

The mandatory English language test was widely debated in the Flemish media, especially in the period following its implementation. Interestingly, the focus was not on the ITACE as a language test, but on the political context in which the test was developed. The analysis is based on media coverage concerning the ITACE in the Flemish press, covering the period before, during and after the test was implemented. Some 30 articles have been analysed qualitatively, drawing on discourse analysis and language policy research (see Barakos & Unger, 2016). The main issues were identified, with a focus on how the ITACE was covered and (mis)interpreted in the media by politicians, lecturers, and language experts. In this section, three of the most salient issues will be briefly discussed: the perceived purpose of the test, its relevance, and its validity.

### 3.1 Purpose

The ITACE was perceived as a political tool of the Flemish nationalists to curb the use of languages other than Dutch. Although the increasing use of English in today's Flemish universities can hardly be compared to the dominance of French in the previous century, the promotion of Dutch and the introduction of English as a medium of instruction remain controversial (van Splunder, 2014, p. 233). Ironically, the 2012 language law, as a result of which the test was developed, was implemented by a Socialist Education Secretary, and not by the Flemish nationalists. Yet, the Government was accused of ‘narrow-minded Flemish nationalism’. Moreover, the ITACE was widely perceived as a result of government interference and excessive regulation.

### 3.2 Relevance

The ITACE was described as irrelevant in the Flemish media. It was argued that the test was not necessary, and that it was humiliating, as a result of which some lecturers refused to take the test. It was also argued that the C1-level was too difficult, even though C1 is the level generally required for academic purposes. This may be based on a misinterpretation of the CEFR and the conceptualization in the media of C1 as the “highest level below the native speaker”. One of the most devastating claims was that “a Professor of Mechanics had to fill in wordlists of birds and amphibians” (De Standaard, 21 February 2015). An equally absurd claim was that the test takers had to produce “Shakespearian English”. Nevertheless, these claims were taken for granted and widely disseminated in the media.

### 3.3 Validity

In spite of globalization and the internationalization of education, language testing is still very much organised according to ‘national’ boundaries. The Association of Language Testers in Europe (ALTE) states that its members can only provide examinations of the official language or languages which are spoken in their own country or region (ALTE Constitution, 2012, article 1). As a result, it was not possible for the ITACE to obtain the ALTE Q Mark (a quality indicator), as English is not an official language in Flanders. In order to overcome the lack of validation by ALTE, the ITACE was validated by an international audit commissioned by the Flemish Government. The audit was based on ALTE’s 17 minimum standards, which have been established as quality profiles for testing examinations. In spite of the fact that the audit stated that the ITACE meets international standards regarding validity and reliability, the positive audit was largely neglected in the media. One article in particular delivered a potentially devastating blow to the ITACE by running the defiant headline “Language Test Fails Test” (De Standaard, 10 November 2014). Even though the article did not support its claim, the tone was set.

## 4 Discussion

The introduction and implementation of the ITACE revealed conflicting discourses regarding language testing. Whereas the Testing Consortium’s concern related to the content of the test, the Flemish Government and the media focused almost exclusively on the context of the test. The Government’s decision to impose a mandatory language test was seized by the Consortium as an opportunity to introduce an interuniversity standardized and validated test, which was developed to assess and to improve the quality of English-Medium Instruction in Flemish higher education. Without this political trigger, it seems unlikely the ITACE would have been developed. It should be noted, however, that the Government was not interested in the ITACE (or any other test) as such. That is, for the Government the language test served a political purpose only, as it was a mere consequence of the Government’s decision to impose a

C1-level which then had to be tested. The media focused on the Government's language policy and the assumed use of the ITACE as a political tool. Hardly any attention was paid to the ITACE as a language test and its possible merits to test and improve the lecturers' level of English as well as the quality of education. This may be due to the fact that the debate was dominated by outsiders – mainly politicians and academics defending their own interests – while the insiders – the Consortium and its language testers – remained remarkably silent. They remained silent as they felt the complexity of the issue could not be conveyed in the media. Moreover, they did not want to be muddled up in politics.

## 5 Conclusion

The ITACE was developed as a language test with a specific purpose: to test whether lecturers in Flemish higher education have a C1-level in English. The ITACE is a standardized and validated test which meets international standards. Ironically, the testing market is mainly organized along national boundaries to protect one's own testing industry. As a result of this protectionism, the ITACE could not be validated by the ALTE. Yet, the ITACE can been seen as an alternative to international tests, which fits in with the idea of English as a lingua franca which is owned by an international community of practice rather than a relatively small group of native speakers which sets the rules and develops language tests. Last but not least, the ITACE made possible and even encouraged collaboration between Flemish universities, bringing together their expertise in language testing. In spite of all this, the ITACE was caught in a media storm, and it was widely perceived as a tool in the hands of the Flemish Government to pursue its language policy. Thus the political circumstances which created the opportunity to develop a language test also discredited the test, and hardly any attention was being paid to the test as a language test.

In order for a language test to be successful, two recommendations may be considered. First of all, acceptance from below is needed. This was obviously not the case in Flanders, where mandatory testing was imposed by the Government, as a result of which the test was seen as a way to control people rather than to provide them with new opportunities. This brings us to the second recommendation: a test should be seen as a door opener rather than a gatekeeper. Even though any test inevitably implies some kind of gate keeping, a test can also be seen as an opportunity. For instance, passing a language test is regarded as a proof of one's command of a language, which may lead to new opportunities (e.g. finding a better job). In order to make a language test more acceptable to test takers, it may be important to stress this empowerment facility of a language test.

## References

ALTE. (2012). *Constitution for the Association of Language Testers in Europe (ALTE)*. Retrieved from http://www.alte.org/docs/constitution-2012.pdf

Bachman, L. F., & Purpura, J. E. (2008). Language assessments: Gate-keepers or gate-openers? In B. Spolsky & F. M. Hult (Eds.), *The handbook of educational linguistics* (pp. 456–468). Oxford, UK: Blackwell.

Barakos, E. (2016). Language policy and critical discourse studies: Towards a combined approach. In: E. Barakos & J. W. Unger (Eds.), *Discursive approaches to language policy* (pp. 23–49). London, UK: Palgrave Macmillan.

Barakos, E., & Unger, J.W. (Eds.). (2016). *Discursive approaches to language policy*. London, UK: Palgrave Macmillan.

Block, D., Gray, J., & Holborow, M. (2012). *Neoliberalism and applied linguistics*. London, UK: Routledge.

Colpaert, J. (2014, November 10). De taaltest voor proffen is zelf gebuisd. *De Standaard*, pp. 38-39.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press. Retrieved from http://www.coe.int/en/web/common-european-framework-reference-languages

De Standaard. (2015, February 21). Taaltest Engels blijft sommige professoren parten spelen. Retrieved from http://www.standaard.be/cnt/dmf20150220_01540974

Flemish Parliament. (2012). *Decree concerning language regulation in higher education*. Retrieved from http://docs.vlaamsparlement.be/docs/stukken/2011-2012/g1655-6.pdf

Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics, 29*, 3–20.

Jenkins, J. (2007). *English as a lingua franca. Attitudes and identity*. Oxford, UK: Oxford University Press.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, UK: Blackwell.

Seidlhofer, B. (2011). *Understanding English as a lingua franca*. Oxford, UK: Oxford University Press.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London, UK: Longman.

Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches*. Abingdon, UK: Routledge.

Spolsky, B. (2012). Language testing and language management. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 495–505). London, UK: Routledge.

van Splunder, F. (2014). Negotiating multilingualism in Flemish higher education. In J. W. Unger, M. Krzyżanowski, & R. Wodak (Eds.), *Multilingual encounters in Europe's institutional spaces* (pp. 221–242). London, UK: Bloomsbury.

van Splunder, F. (2016). Language ideologies regarding English-medium instruction in European higher education. Insights from Flanders and Finland. In: E. Barakos & J. W. Unger (Eds.), *Discursive approaches to language policy* (pp. 205–230). London, UK: Palgrave Macmillan.

Wächter, B. (2014). *The European map of English-taught programmes in 2014: Results of a new ACA study*. Paper presented at the ACA, Brussels, 4 December 2014.

Widdowson, H. G. (1994). The ownership of English. *TESOL Quarterly 28*(2), 377–89.

# Empowering Learners for a Demanding Labor Market: The "Groups for Experimentation of Plurilingualism" Education Program in Catalonia

**Montserrat Montagut Montagut**, Ministry of Education (Catalan Government – Generalitat de Catalunya), Spain

**Abstract:** Mastering the technical skills of a job is no longer enough for a demanding and highly competitive job market. Education systems across Europe have echoed this reality and Catalonia hasn't been an exception. In recent years, the Department of Education has launched several innovative programs in order to bring about deep changes in the educational system, both pedagogical and organizational. One of these programs is the "Plurilingual Generation" program, formerly known as the "Groups for Experimentation of Plurilingualism" program (GEP). This innovative educational program was launched in 2013 with the aim of activating the plurilingualism of students in at least three languages by developing interdisciplinary and transversal school projects. The program also reinforces professional skills and prepares students for lifelong learning. And it does so with a double aim: to facilitate the entering of students in the labor market and to promote a responsible and active citizenship. More than 400 Catalan schools have participated in it since its beginning.

## 1 New jobs and skills for a global world

Since the second half of the 20th century and partly as a result of globalization, our societies have strongly changed. Changes in economy, culture and technology have contributed to this deep transformation. On the other hand, the economic crisis in Europe revealed the fragility of the system, caused the loss of thousands of jobs and threatened social cohesion. As a result, the gap between the actual skills of workers and skills that the market needs became apparent.

Improvements in technology and also changes in everyday interactions have created a need for new types of jobs and occupations. Advanced robotics, artificial intelligence, augmented reality, cloud computing, etc. are just some examples of new jobs. Mastering the technical skills of a job is no longer enough for a demanding and highly competitive job market. Professionals must be able to analyse reality from a critical point of view, solve problems in a creative way and be used to teamwork. And one of the most valued skills is communicative competence in general and foreign languages proficiency in particular.

The 21st century labor market carries with it new kinds of tasks. Nowadays occupations require non-routine tasks. Success in the labor market depends on the worker's capacity to develop tasks which have a strong communicative component, which are very demanding from the cognitive point of view and which imply good capacity for interacting with those from different cultures and backgrounds.

So, which are the skills needed for this new landscape? In many industries and countries, the most in-demand occupations did not exist 10 or even five years ago. According to the World Economic Forum (2016) report on the future of jobs, skills like complex problem solving, critical thinking, creativity, etc. are crucial. Overall, social skills – such as persuasion, emotional intelligence and teaching others – will be in higher demand across industries than narrow technical skills, such as programming or equipment operation and control. In essence, technical skills will need to be supplemented with strong communicative, social and collaborative skills.

With labor markets and the demand for skills changing, education systems need to adapt. However, despite widespread investment, education systems continue to fall short in providing the right skills for employability.

## 2 Improving foreign languages proficiency for better economic results

To overcome this situation the European Commission has launched several strategies in order to meet growing demand for higher skills levels and reduce unemployment. One of its best known strategies is the communication *Rethinking Education: Investing in skills for better socio-economic outcomes* (European Commission, 2012). The initiative focuses on three areas in need of reform: quality of education, accessibility and funding. And it considers that reforms should be designed to raise basic skills levels, to promote apprenticeships and entrepreneurial skills and to improve foreign language skills. The last point is of special interest for the European Commission as foreign language proficiency is considered to be one of the main determinants of professional mobility, employability and personal development of European citizens, in particular young people, in line with the objectives of the Europe 2020 strategy for growth and jobs.

Just a year before, the European Commission had published a report providing a compendium of good practices and guidelines for modernizing education systems and reducing the mismatch between supply and demand of language competences. The document insists on the importance of raising the general level of language competences, broadening the range of languages taught, re-orienting teaching contents towards professional purposes, and improving the training of staff.

On the other hand, the 2011–12 Survey on Language Skills held in 14 European countries revealed an insufficient proficiency of foreign languages among pupils at the end of compulsory education, and a more recent report confirms that foreign language teaching in many countries remains inadequate. In 2014 the Council of the European Union invited the member states to adopt and improve measures aimed at promoting multilingualism and enhancing the quality and efficiency of language learning and teaching, including by teaching at least two languages in addition to the main language(s) of instruction from an early age and by exploring the potential of innovative approaches to the development of language competences. Following this recommendation and in order to support EU member states' strategy, the European Commission published the *Improving the Effectiveness of Language Learning* report (2014), which focuses on two scientifically proven methods of speeding up language learning: Content and Language Integrated Learning (CLIL) and Computer Aided Language Learning.

All education systems across Europe have echoed this reality and its consequences. Many countries in Europe have taken into account the institutional recommendations and are doing their best to improve their education systems and to increase the knowledge of foreign languages among their students at all levels (primary, secondary and post-secondary education). Catalonia hasn't been an exception.

## 3 Plurilingual education: the Catalan strategy for the new generations' success

The education policies in Catalonia focus on two aspects which are at the true core of the system: competencies and inclusion. That means that the basic goals are: providing citizens in new generations with the skills they need for their professional success and in order to be able to face future unexpected challenges and, on the other hand, reinforcing social cohesion, ensuring everybody's access to education and highlighting diversity (which is an added value even for competitiveness).

In recent years, the Department of Education (Ministry of Education in Catalonia) has launched several innovative programs in order to foster changes, both pedagogical and organizational, to ensure a better match between skills acquired at school and skills required in real life. One of these innovative programs is the "Plurilingual Generation (GEP)" program, formerly known as "Groups for Experimentation of Plurilingualism". This program is part of a broader governmental strategy aimed at improving our students' linguistic and communicative competence through the implementation of an education model based on plurilingual education. This strategy follows and adapts to the recommendations of the Language Policy Unit of the Council of Europe on Plurilingual and Intercultural Education.

There are several methods for implementing a plurilingual education. These approaches complement the individual methods used to teach each language. The two most complex approaches are content and language integrated learning (here language can be both the language of instruction and a foreign language) and integrated learning of languages. According to the Council of Europe, these two complex approaches are key for the right implementation of a plurilingual education. With CLIL, languages are learned through different subjects and, in some way, all teachers become language teachers. We develop this approach through several programs, one of which is thePlurilingual Generation program.

The integrated learning of languages proposes collaborative work among teachers of different languages. This approach focuses on the common features of languages and claims the benefit of helping students to use their linguistic knowledge and whole repertoire to learn other languages. Last year we launched a project to this end called "Go ahead: Integrated teaching and learning of languages". Other proposals, such as awareness-raising activities on linguistic diversity and openness to other cultures, intercomprehension activities among related languages, real or virtual mobility actions, etc. are considered complementary to the aforementioned methods. The Department of Education in Catalonia also fosters these types of activities.

The "Plurilingual Generation (GEP)" program follows the recommendations of the European Commission in general and, in particular, those included in the 2014 European Commission  document. This program was launched in 2013 as an experimental program. Since then it has evolved, changed and been improved.

## 4 The Plurilingual Generation Program (GEP)

### 4.1. General objectives

The general objectives of the program are, first of all, increasing students' time exposure to the foreign language (as this language has little presence outside school); secondly, improving our students' communicative competence in a foreign language (at this moment, the program focuses on English and French); and finally, helping students to acquire 21st century skills and lifelong learning strategies. However, with this initiative we would also like to foster the deployment of interdisciplinary school projects based on the project based learning (PBL) approach and in the efficient use of ICT, encouraging team teaching (collaboration between subject teacher and language specialist teacher), assisting head-teachers in introducing improvements in school management and curriculum design, and promoting dissemination and exchange of good teaching practices among schools.

We have chosen this double approach (CLIL and PBL) because we believe it is the perfect combination to provide meaningful content and context for focus on issues that 21st century students are facing or will face in their work, their lives and their future (critical thinking and problem solving), to offer opportunities to negotiate meaning and communicate in relevant and authentic ways using a range of media (communication) and to engage learners in working together using a variety of resources and texts, including the Internet to develop knowledge and skills (collaboration).

As stated at the start of this paper, critical thinking and problem solving, communication and collaboration are basic skills for success in the labor market. Moreover, content brings the real world into the class and provides a means for developing more advanced language proficiency and transferring this to academic contexts.

Some of the ideas and concepts behind CLIL have been present in Catalonia for over 20 years. The first institutional programs aimed at promoting language learning projects, content teaching and CLIL projects go back to 1994. And many teachers started learning about foreign language learning integrated approaches, mainly through language learning activities in the first Erasmus program in 1987, the subsequent Socrates I and II and the Lifelong Learning Programme, which is the predecessor of the current Erasmus+ program.

With this new program we promote new organizational structures within schools (as the steering committee). Also, participating schools have to design a strategic action plan of the project deployment and the teachers involved need to adapt the project to their school year program in advance. This innovation program has strong and intensive support from the Education Inspectorate in order to control the project deployment within each school and its assessment.

### 4.2. Participation requirements and implementation of the program within a school

Only public funded schools can participate in this program, which lasts for three school years, and they need the prior approval of teaching staff and also of the School Council, which is part of the school community within a school's management (the members are representatives of the teaching staff, pupils, families, town council, etc.). The schools can be primary, secondary and post-secondary schools (including vocational education and training) and participation is subject to compliance with some prerequisites: for instance, the school education project has to include a specific aim relating to plurilingual education, and the teaching staff who will participate should have a minimum knowledge of the foreign language (nowadays the minimum is B2, but we give priority to teachers with a C1 or a C2 certificate).

But participation also means a strong commitment to the program, as the school needs to warrant:

(1) The constitution of a steering team in order to control and provide internal assessment of the project. The members are: the headmaster, the foreign language teacher specialist (in the case of secondary education, the head of the school foreign language department), participants in the training sessions. Also part of this team can be other teaching staff who, despite not being involved in the training sessions, have received the training and are actually implementing the program with some of their groups of students.

(2) The design and implementation of an Action Plan: this is an internal working document aimed at facilitating the monitoring of the school project deployment and the evaluation of the final outcomes. It has to include: the diagnosis of the initial situation (sociolinguistic school context, students' and teachers' language competences, etc.); the objectives (courses and subjects involved); the timing for a progressive deployment of the new approach; the expected outcomes on students' competence and on school organization, teachers involved, etc.

(3) The active participation in the training activities and presentation of a final activity report which will be evaluated by the Education Inspectorate Body. In cases where this evaluation is positive the teachers are awarded with a certificate of innovation, which is of high value in a teaching career.

### 4.3. Program training activities

During these three years of participation the directive staff receive 10 hours of training, consisting of advice and tools in order to design a good action plan and to guarantee the correct monitoring of the whole project; and the teachers (a minimum of two and a maximum of three per school) will receive 90 hours of training split into two blended-learning courses focused on the development of a CLIL unit based on PBL. In this program the majority of the participants are teachers of subjects such as mathematics, natural and social sciences, chemistry, history, physics, etc. In every school only one teacher of foreign languages can participate, and team teaching is promoted. The training for teachers helps them to develop a project based on PBL and CLIL; this project has to incorporate some use of technology in an effective and sensitive way and can incorporate some intercomprehension activities (especially in the French groups) as a way to foster the students' plurilingual competence. We also offer some complementary workshops to provide teachers with some strategies to help their students develop their linguistic competence.

In the next school year (2017–2018) all the trainers will come from the education system. That means that they will be civil servants who are actually teaching in a school and who have been previously trained to become trainers of this program. But the program is also a good example of cooperation with other institutions involved in education and in the teaching of foreign languages. Two universities are helping us with the deployment of this program: Autonomous University and the University of Barcelona on the one hand, both involved in the basic training activities and, on the other hand, the British Council, Oxford University Press, Trinity College, Cambridge University Press and International House, all of them involved in the complementary training activities.

## 4 Conclusion

The experience began in 2013 with 53 schools. Every year we now expect to see a minimum of 100 new schools. From 2013 to 2016, 383 schools participated and 600 teachers were trained. These figures represent almost 10% of schools in Catalonia. In five years' time (2022) we will see the global results and impact of the program as the Catalan Ministry of Education is going to evaluate it. The objective of this assessment will be to check the effectiveness of the GEP program in fulfilling the basic objective of this strategy, which is: to provide students with a solid communicative and plurilingual competence that contributes to their academic growth and subsequent job placement and that enables them to interact with a global world in a critical way.

## Further reading

Council of Europe. (2007). *From linguistic diversity to plurilingual education. Guide for the development of language education policies in Europe*. Strasbourg, France: Language Policy Division, Council of Europe.

Departament d'Ensenyament de la Generalitat de Catalunya. (2013). *L'escola catalana: un marc per al plurilingüisme*. Retrieved from http://xtec.gencat.cat/ca/projectes/plurilinguisme/sobre/presentacio

Departament d'Ensenyament de la Generalitat de Catalunya. (2013). *Ofensiva de país a favor de l'èxit escolar. Pla per a la reducció del fracàs escolar a Catalunya 2012–18*. Retrieved from http://ensenyament.gencat.cat/web/.content/home/departament/publicacions/monografies/ofensiva-exit-escolar/ofensiva_exit_escolar.pdf

Govern de la Generalitat de Catalunya. *Pla de Govern de la XI Legislatura*. Retrieved from http://www.govern.cat/pres_gov/estilos/govern/pdf/Pla_de_Govern_XI_legislatura.pdf

Pereña, M. (coord.) (2016). Ensenyar i aprendre llengües en un model educatiu plurilingüe. Metodologies i estratègies per al desenvolupament de projectes educatius i per a la pràctica docent. *Quaderns d'Educació, 75*. ICE-Horsori. Universitat de Barcelona.

## References

Beacco, J. C., Byram, M., Cavalli, M., Coste, D., Egli Cuenat, M., Goullier, F., & Panthier, J. (2016). *Plurilingual and intercultural education. Guide for the development and implementation of curricula for plurilingual and intercultural education*. Strasbourg, France: Council of Europe.

European Commission. (2012). *Rethinking Education: Investing in skills for better socio-economic outcomes*. Retrieved from www.cedefop.europa.eu/files/com669_en.pdf

European Commission. (2014). *Improving the effectiveness of language learning: CLIL and computer assisted language learning*. Retrieved from http://ec.europa.eu/dgs/education_culture/repository/languages/library/studies/clil-call_en.pdf

World Economic Forum. (2016). *The Future of Jobs*. Retrieved from http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

# Hong Kong Diploma of Secondary Education (HKDSE) English Language Level Descriptors: Stakeholder Recognition and Understanding

**Neil Drave**, Assessment Development Division, Hong Kong Examinations and Assessment Authority, Hong Kong

**Abstract:** This paper summarises research conducted by the Hong Kong Examinations and Assessment Authority (HKEAA) on the English Language Level Descriptors (LD) for Hong Kong Diploma of Secondary Education (HKDSE). The aim of the research was to find out whether the LD for the Writing and Speaking paper were understandable and meaningful to stakeholders. Data were gathered using a 90-item questionnaire, an evaluation of samples of marked candidate work, marking of candidates' work, and interviews. It was found that the stakeholder groups considered the LD to be fit for purpose, in general. The stakeholders were generally able to understand the LD and to use them to rate candidate performances accurately. There were some difficulties in understanding specific components, some were felt to be redundant and there were suggestions for adding others. In the opinion of the Steering Group which oversaw it, the research has given interesting insights into the opinions of stakeholders and suggested useful actions to be taken to improve the LD. Progress reports on the research were presented at the Academic Forum on English Language Testing in Asia (AFELTA) 2015 (Drave, 2015) and 2016 (Drave, Shiu and Chan, 2016).

## 1 Introduction

The Hong Kong Examinations Authority (HKEAA) employs standards-referenced reporting (SRR) of results in Hong Kong Diploma of Secondary Education (HKDSE) examination. This means that 'candidates' levels of performance are reported with reference to a set of explicit and fixed standards of performance' (HKEAA, 2011). In SRR, the level descriptors (LD) used to describe performance are an important link between the assessment body and the stakeholders who use them so they must fulfil certain criteria for usefulness.

In the LD three traits ('domains') are assessed at each level: Content, Language and Organisation. HKDSE practice is to sum the domain scores to derive an overall score in each paper and then convert the scores to levels in a separate grading process. Although there are LD for all papers, it was decided to focus on the Writing and Speaking LD in this research.

## 2 Research objectives

The research set out to investigate the following questions:

(1) Are the LD relevant to the work of stakeholders?
(2) Are the LD understandable to stakeholders?
(3) Are the LD (perceived to be) meaningful?
(4) Are the LD (perceived to be) complete?
(5) Do stakeholders think that the LD match the candidates' performance in writing/speaking?
(6) Are different groups of participants equally able to understand and use the LD?

## 3 Participants

Participants (n=49) were grouped as follows:

- Group 1: Trained Experts (TE) (n=10)

These were teachers who had been markers of HKDSE English Language Writing paper (Group 1a, n=5) or Speaking paper (1b, n=5). These participants were likely to be familiar with the LD.

- Group 2: Untrained Experts (UE) (n=20)
  These were English language teachers who had not previously marked HKDSE from schools (2a, n=8) or tertiary institution language centres (2b, n=12). They were familiar with rating scales, but not necessarily the LD.

- Group 3: Expert Users (EU) (n=7)
  These were university/college personnel who might be using LD to make admissions or streaming decisions.

- Group 4: Non-expert Users (NEU) (n=12)

These were administrative or Human Resources (HR) personnel from non-education fields who made appointments and similar decisions but were probably not using the LD to do so.

## 4 Methodology

Participants were recruited using the HKEAA's marker database. A 90-item online questionnaire containing Likert-type and open-ended items was administered to them.

Research Days were held, in which participants undertook three activities:

- Scoring of Speaking recordings: 24 candidates (in six groups) from the 2015 HKDSE English Language Speaking paper were shown to participants.
- Scoring of Writing scripts: 25 Writing scripts from the 2015 HKDSE English Language Writing Paper were randomly distributed to participants.
- Interviews: participants were interviewed.

An interactive online questionnaire was designed. Through the questionnaire, respondents evaluated two samples of performance at each level in Writing and eight Speaking candidates. The performances had been previously marked and participants were asked to what extent they agreed with these scores.

A testing instrument was designed for Speaking. This was a score sheet in which participants recorded scores for the 24 Speaking candidates during the Research Day.

An interview pro forma was designed. This 10-question document was used to structure interviews.

There were both quantitative and qualitative approaches to data analysis. Qualitative analysis involved iterative coding of the interview transcriptions until ‘saturation point’ (i.e. the point at which no new information emerged). Responses to open-ended questions were summarised.

## 5 Findings: questionnaire

### 5.1 Questionnaire findings: Part 1

The NEUs were less familiar than other groups with HKDSE and the LD. The majority said that the LD were useful and found them easy to understand.

### 5.2 Part 2 – Understandability, usefulness and detail

Participants were asked to give their opinion on whether the LD at each level were easy to understand, useful for describing language proficiency and detailed enough. Their responses were Likert-type items on a 6-point scale, where 1 indicated ‘Strongly disagree’ and 6 ‘Strongly agree’.

#### *5.2.1 Writing*

Respondents felt that the LD Writing as a whole were clear, useful and detailed enough. It was felt that Levels 4 and 5 were clearer than others. Almost all respondents felt that there was a clear progression, consistency from one level to another and enough levels. The NEU were the least positive about all of these aspects. Level 1 was the least favourably received.

#### *5.2.2 Speaking*

Respondents felt that the LD Writing as a whole were clear, useful and detailed enough, with Levels 4 and 5 felt to be clearer than others. Almost all respondents felt that there was a clear progression, consistency from one level to another and enough levels. The NEU were the least positive about all of these aspects, with the lowest mean scores on all questions. Level 1 was the least favourably received.

### 5.3 Part 3 – Open-ended responses

The questionnaire contained open-ended elements in the form of statement prompts which probed participants’ views on different aspects of each level in the Writing and Speaking LD.

The following conclusions can be drawn:

- Participants needed more help with understanding the LD in the form of examples or glossaries with definitions.
- Certain lexical items seem to have caused problems of understanding.
- Participants encountered the same/similar wording in different places, such as different levels, but were unsure about whether they meant the same thing.

- Certain criteria occurred in some levels but not others, and participants were unsure of what to make of this.

## 6 Findings: evaluation of samples using the LD

### 6.1 Writing

The majority of participants in all groups judged the candidates' performance to be 'About the same standard' or 'Exactly the same standard' as the actual score it had been assigned at all levels, except for Level 4 (Exemplar 2), where half the respondents (49%) felt that it was 'Worse' and 6% that it was 'Much worse'. There was also some disagreement at Level 3 (Exemplar 2), with about a fifth of respondents (18%) feeling that the performance was 'Worse' than the descriptor suggested and 10% feeling that it was 'Better'.

There was no significant difference between the groups in their judgements, except at Level 4, where EU tended to give lower scores. The (non-statistically significant) general trend was for participants to judge the exemplars as being worse than the descriptor suggested. At Level 5 and Level 3, only TE thought the exemplars and descriptors were congruent, while all other groups felt they were worse. At Level 1, only NEU thought the two were congruent, while all other groups thought the exemplars were better.

It seems that participants expected a higher standard of performance than they were given to review, except at the lowest level, when they expected worse performance.

### 6.2 Speaking

The majority of participants judged the candidates' performance to be 'About the same standard' or 'Exactly the same standard' as the actual score it had been assigned at all levels.

There was no difference between the groups in their judgements, except at Level 4, where UE and NEU tended to give higher scores than EU, meaning that they felt the performance warranted a higher level than had been awarded.

### 6.3 Summary

There was some variability in the ability to evaluate the samples, and this cut across groups. Participants seemed to be more accurate at scoring Speaking; in Writing, they seemed to have expected to see better work. In general, there is evidence that the LD do match the performance seen in the samples, with the possible exception of Level 3 Writing.

There are two caveats to note. It is possible that the exemplars did not accurately represent the putative standards of the different levels. Also, it is possible that inaccuracies in assigning levels by EU were due to a lack of correspondence between the LD and the marking scheme, which some participants would usually use, rather than any deficiencies in the LD *per se*.

# 7 Findings: marking of candidate work

Participants assigned levels to Writing and Speaking samples. If the ratings are described as ‘accurate’, the participant scores (i.e. levels) matched the official scores.

## 7.1 By group

### *7.1.1 Writing*

The ratings were too generous overall:

- More than half the candidates were scored too highly (one level) by all groups.
- Nine candidates were scored too harshly, and this included all five of the Level 5s.

Group differences:

- EU were most accurate at Levels 1 and 2 but least accurate at Levels 4 and 5.
- UE were most accurate at Level 4 while TE were most accurate at Level 5.
- Overall, NEU were the least accurate. The group differences were not statistically significant.

These findings contrast with those for the evaluation phase, in which participants were harsh.

### *7.1.2 Speaking*

In general, the scores were accurate, with the following exceptions:

- One candidate was scored too highly (one level) by all groups (actually Level 1 but the majority gave Level 2).
- One candidate was scored too highly (one level) by all groups except EU (actually Level 2 but given Level 3).
- One candidate was scored too low (one level) by all groups (actually Level 3 but given Level 2).

Group differences:

- UE gave high scores to 3 candidates. All these actually scored Level 4 but were given Level 5 by about half, a third and a fifth of UE participants, respectively.
- NEU and TE were most accurate in their scoring overall, while EU were least accurate, and this difference was statistically significant.

### *7.1.3 Summary*

Assuming that the chosen exemplars were accurate reflections of the intended standards, one might conclude that the LD were usable for Writing since all the groups were equally accurate in their scoring. The possible exception is Level 5, which was underused. For Speaking, one can draw the same conclusion, but for a different reason: in this case, the EU were the least accurate, but the NEU were as accurate as the TE, which suggests that specialist training is not required to give accurate judgements.

## 7.2 By marker

### *7.2.1 Writing*

The correlation between all the participants' scores and the actual scores was 0.96, which is very high.

There is evidence of a 'central tendency' effect (Knoch, 2009) in Writing, however, with almost all markers overrating at Level 1 (89.6%), close to two-thirds overrating at Level 2 (60.4%) and more than half overrating at Level 3 (52.1%). In contrast, more than three-quarters of participants (79.1%) underrated at Level 5.

### *7.2.2 Speaking*

The markers' scoring was accurate on the whole, to within one level. Only seven markers were out by more than one level. The correlation between all participants' and actual scores was high (0.96). Participants' scores mimic the actual scores and all markers rank the candidates similarly.

As for Writing, there is evidence of a central tendency, with markers overrating at Levels 1 (68.8%) and 2 (85.4%) and underrating at Level 4 (68.8%). One marker did not give any Level 1s.

### *7.2.3 Summary*

- Rating of both Writing and Speaking was satisfactory in general, as judged by correlation figures and data on overrating and underrating.
- There was a 'central tendency' effect, especially for Writing.
- There were few group differences in ratings, but some individual ones.

# 8 Findings: interview data

- There was a high degree of overall satisfaction with the LD.
- There were few differences between the stakeholder groups in how they evaluated the LD and whether they considered the LD at different levels to be useful, detailed, clear enough and consistent. The small number of participants in some of the groups means that any conclusions are speculative, however.

- Some interviewees felt that the standards instantiated by the LD were too low, particularly at the top of the scale.
- Some wanted negative aspects of performance to be included in the LD at the lower levels, i.e. what candidates cannot do.
- Some requested more help in the form of examples or glossaries with definitions.
- There were comments on the (lack of) consistency of terms/concepts from one level to another and in different domains e.g. certain features occur in some levels but not others.
- Some disagreed with some of the features included in the LD, e.g. ‘body language’ and ‘prompting’ in Speaking and ‘creativity and imagination’ in Writing.
- There were concerns about lexical items with an unclear scope or frame of reference, e.g. ‘ambitious’, ‘familiar’ and ‘simple’.
- Many comments pointed to words and phrase which were vague, especially relating to non-numerical vague quantifiers (Channell, 1994; e.g. ‘some’) and inherently vague lexis which implied a particular amount of knowledge or skill (such as ‘range’).

## 9 Overall findings

In this section, findings are presented according to the research questions.

(1) Are the LD relevant to the work of the stakeholders?

The degree of familiarity with the LD depended on whether or not participants had worked with the HKEAA. The HKDSE qualification was not universally known, and there was some confusion about the relationship between LD and marking documents.

(2) Are the LD understandable to stakeholders?

Participants felt that they could make sense of the LD in a general way and that the descriptions were understandable without specialised knowledge. The expert users were confident that they were interpreting the terms correctly. They pointed out the inherent vagueness and subjectivity of some language items (e.g. non-numerical quantifiers); they also recognised that these are an inherent feature of descriptors and may not be amenable to improvement, however.

(3) Are the LD (perceived to be) meaningful?

LD do instantiate important features of English performance (language, non-language). Participants expressed reservations about the relevance of body language and creativity, however, since these did not seem to be demanded in the tasks the candidates were given and were also difficult to interpret.

(4) Are the LD (perceived to be) complete?

Participants were in general satisfied with the completeness of the LD. Some felt that certain features were not captured, there was confusion about the weighting of features within the LD, and requests for additional materials e.g. examples.

(5) Do stakeholders think that the LD match the candidates' performance in writing/speaking? How close is the match perceived to be?

Most participants agreed that the official levels were the 'correct' ones. Most participants were able to match the candidates with the correct levels, but many were unwilling to give 5s and 1s in Writing. This may be because of unfamiliarity with the standards rather than the quality of the LD.

(6) Are the different groups of participants equally able to understand and use the LD?

It seems that the intra-group variation was more important than the inter-group, or at least there were no consistent patterns in this regard.

## 10 Conclusion

This research has partly validated the utility of the LD as public relations and accountability documents. At the time of writing, it is likely that there will be some major changes to the English Language Writing paper, meaning that any possible amendments to the LD will be considered in tandem with these proposed changes. As for Speaking, the Steering Group which oversaw the project felt that the research had served to highlight the limitations of the current Speaking paper, which could be reviewed. In the meantime, the researchers were authorised to amend the Speaking LD as necessary.

## Further reading

Details of the HKDSE English Language can be found at http://www.hkeaa.edu.hk/en/hkdse/assessment/subject_information/category_a_subjects

In HKDSE English Language, there is a set of LD for each subject and for each paper, which are posted here: http://www.hkeaa.edu.hk/en/hkdse/assessment/subject_information/category_a_subjects/hkdse_subj.html?A1&1&2_4

Samples of candidate work at each level are also posted on the HKEAA web site.

For details of historical developments in Hong Kong public examinations before the implementation of the DSE, see Choi, C.-c. & Lee, C. (2010). Developments of English Language assessment in public examinations in Hong Kong, in L. Cheng, & A. Curtis (Eds.), *English Language Assessment and the Chinese Learner* (pp. 60–76). London: Routledge.

## References

Channell, J. (1994). *Vague Language*. Oxford: Oxford University Press.

Drave, N. (2015). Hong Kong Diploma of Secondary Education (HKDSE) Level Descriptors: Usability and meaningfulness. *Proceedings of the 2015 AFELTA* (pp. 51–68). Tokyo: Japan.

Drave, N., Shiu, J., & Chan, A. (2016). Research on Hong Kong’s Diploma of Secondary Education (HKDSE) English Language Level Descriptors: Progress report, in *Proceedings of the 2016 AFELTA. Hong Kong* (pp. 125-155). Retreived from: http://www.hkeaa.edu.hk/DocLibrary/Event/AFELTA_2016_Proceedings.pdf

HKEAA. (2011). *Grading Procedures and Standards-referenced Reporting in the HKDSE Examination*. Hong Kong: HKEAA.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. Frankfurt: Peter Lang.

# Academic Literacy and Language Proficiency in Testing: Overlapping and Diverging Constructs

**Kevin Cheung**, Cambridge English Language Assessment, United Kingdom
**Mark Elliott**, Cambridge English Language Assessment, United Kingdom

**Abstract:** The construct of language proficiency focuses on speakers of other languages, whereas academic literacy (AL) is conceptualised as a set of skills needed for successful university study, which many native speakers struggle to develop. Clearly, there is overlap and divergence between the two and both inform English for Academic Purposes (EAP) assessment. By employing a sociocognitive model of language proficiency (Weir, 2005) as a basis, this paper examines commonalities and distinctions between the two concepts. Assumptions about AL and general language proficiency (GLP) are revealed when these constructs are examined closely in relation to each other.

A number of questions about these constructs are important for language testing researchers and practitioners to consider. Making the questions explicit and posing them to the language testing community potentially informs future EAP assessment, pedagogy and research.

## 1 Introduction

General language proficiency (GLP) and academic literacy (AL) are constructs used widely in pedagogic contexts. They often focus on higher education; developing high levels of GLP is crucial for students intending to study in a territory of their choice, and AL is viewed as important for ensuring success once university study has commenced. Some areas of difference and similarity between the constructs are easily identified; however, they also intersect in ways that have not been examined closely before. The recent trend of globalisation in higher education makes it important for universities to develop efficient forms of support for learners from various cultures. Programmes that improve AL are possibly different to those aiming to improve language proficiency, and this potentially impacts on how assessment can support learning.

Therefore, it is important to compare language proficiency as operationalised by the Common European Framework of Reference for languages (CEFR, Council of Europe, 2001) with AL as defined by universities, to examine how the constructs align theoretically. This forms the focus of the present paper. We briefly introduce models of GLP and AL, before moving onto a comparison of these constructs. Using example tasks and definitions, questions are posed to language testing practitioners, and potential avenues of further research are identified.

## 2 Models of general language proficiency

GLP is typically modelled on one of two distinct but related approaches: cognitive processing models and functional models, which will now be discussed in turn.

### 2.1 A cognitive processing model

A cognitive processing model treats language use as a sequence of hierarchical but interacting processes in which the activation of each higher level of processing requires the activation of all lower levels, interacting in highly complex ways. Such models exist for reading (Khalifa and Weir, 2009), writing (Shaw and Weir, 2007), listening (Field, 2013) and speaking

(Field, 2011). Considering the case of reading, the model can be broken down into six separate stages, as shown in Figure 1.

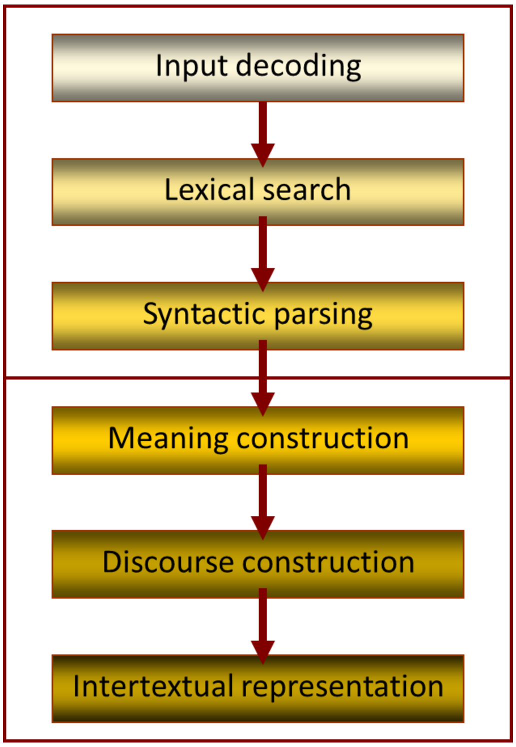


**Figure 1.** A cognitive processing model of reading (adapted from Khalifa and Weir, 2009)

The first three stages, up to syntactic parsing, constitute the lower-level processes, resulting in the establishment of a 'bare proposition' from a sentence – a literal, decontextualised proposition, possibly with multiple interpretations, drawn straightforwardly from the lexis and grammatical structure of the sentence. The next two levels of processing, the higher-level processes, involve enriching the meaning of the bare proposition with contextual cues and the reader's world knowledge, then incorporating the new proposition into a representation of the discourse to date, relating it to previous propositions within a hierarchical structure. The final stage of processing only takes place when the reader is combining meaning from multiple sources to create a unified intertextual representation; this is naturally more complex than working with a single text since the coherence and cohesion which would exist within one text will not exist across texts and there may be conflicting and contradictory information.

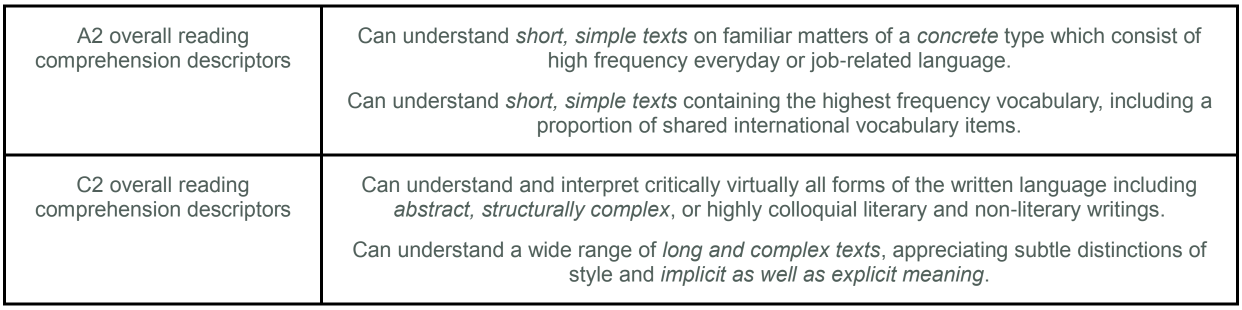
Within a cognitive processing model, progression is defined as expertise: increased efficiency and ultimately automation of successive levels of cognitive processing, freeing up scarce working memory resources in order to access increasingly higher levels of processing which are initially unavailable due to working memory constraints. Such a model could be seen as the more scientifically 'correct' approach in that it can have predictive power.

Intertextual processing, as described above, is of particular interest in terms of the current topic since it is a typical demand of students in higher education when working on assignments of critical writing which involve drawing from multiple sources written from differing perspectives – although it should be noted that it is not solely connected to academic contexts; for example, multiple film reviews may be consulted when planning a trip to the cinema.

### 2.2 A functional model

Unlike a cognitive model, which focuses on latent activity, a functional model describes the output of language processing in terms of the specific tasks a language user is engaged in. Language functions may be considered epiphenomena of cognitive processing in that they represent visible, or at least interrogable, surface activities which result from underlying cognitive processing. The mapping from cognitive processing is not always clean, and as such functional models do not have the same predictive power as cognitive processing models, but they are both intuitive and practical as they can be used to describe expected behaviours of language users in real-world contexts.

The CEFR (Council of Europe, 2001) is the best known example of a functional model of language proficiency, with descriptors describing both the types of task, contexts of use and types of texts language users are able to handle across six proficiency levels, A1 to C2. Returning to the example of reading, descriptors for A2 and C2 for overall reading comprehension (Council of Europe, 2001:69) are (emphasis added):



| | |
|---|---|
| A2 overall reading comprehension descriptors | Can understand *short, simple texts* on familiar matters of a *concrete* type which consist of high frequency everyday or job-related language.<br>Can understand *short, simple texts* containing the highest frequency vocabulary, including a proportion of shared international vocabulary items. |
| C2 overall reading comprehension descriptors | Can understand and interpret critically virtually all forms of the written language including *abstract, structurally complex*, or highly colloquial literary and non-literary writings.<br>Can understand a wide range of *long and complex texts*, appreciating subtle distinctions of style and *implicit as well as explicit meaning*. |

**Table 1.** CEFR descriptors (Council of Europe, 2001:, p. 69, emphasis added)

As emphasised, key distinctions are drawn between the types of texts learners can handle at the two levels in terms of their length, complexity and their nature (abstract versus concrete) – all three of these dimensions are known to increase cognitive demands when processing a text. Interestingly, high level proficiency as defined by the CEFR is determined almost exclusively in this way, with little to no attention paid to the more interpersonal type of language used in everyday relationships, for example, rapport building and telling jokes. A distinction could be made between the handling of what Kay (1977) described as *autonomous language*, following Bernstein's (1964) concept of *elaborated codes*. Autonomous language describes the types of language used by speech communities which collectively own a body of knowledge greater than that which any individual can possibly know; knowledge cannot always be assumed, so relations and contingent information need to be explicitly signalled, resulting in complex texts with considerable use of subordination and complex noun phrases. This contrasts with *non-autonomous language* (analogous to Bernstein's (1964) *restricted codes*), which are employed in tight-knit speech communities such as families where great amounts of knowledge can be assumed and do not need be explicitly related, resulting in less complex texts. The focus of the CEFR at higher levels is strongly on autonomous language, which is precisely the type of language associated with academic (and professional) contexts. The skilful use of non-autonomous language is also a high-level language skill, although of a different nature, and should not be confused with the type of language use exemplified by the lower levels of the CEFR; these could be related to Klein and Purdue's (1997) notion of a *basic variety*, or Hulstijn's (2007) concept of core language proficiency – a simple version of a language with lexis and grammatical complexity suitable for everyday situations without either the complexity associated with autonomous language or the nuances associated with non-autonomous language. In this sense we can claim that there is an inherently academic dimension to high-level language proficiency, at least as conceptualised by the CEFR, which we argue does not cover the entirety of language proficiency, as shown in the tentative model outlined in Figure 2.

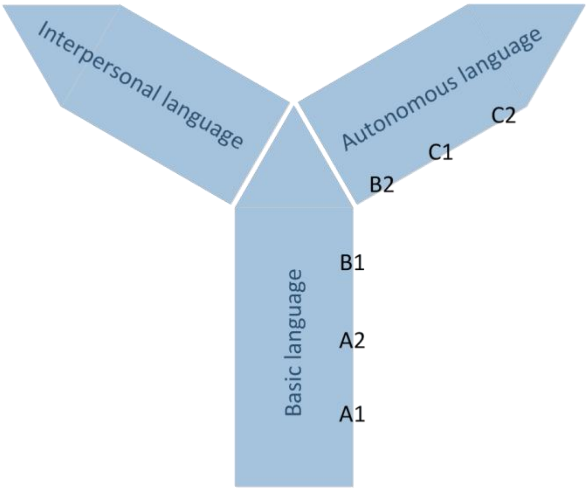


**Figure 2.** Language proficiency and the CEFR (Elliott, 2011)

It should be noted that this the above should not be considered surprising, nor is it intended as a criticism of the CEFR, since the purpose of the document is to facilitate comparisons across languages in teaching and assessment, which is dictated in part by practicality; schools and certificate end users would be likely to have little interest in teaching or testing a learner's joke-telling ability, even in the unlikely event that it could be taught or, even less likely, tested in any meaningful way. Nonetheless, there is a clear relationship between the higher levels of the CEFR, beginning in its nascence at around B2 but more clearly represented at the C levels, and academic language use. This divide could perhaps be seen as a reason that many native speakers of a language may struggle on a C2 level test, since they may not be experienced in the specific, education-related, language uses, while a non-native learner who has studied in their own country and passed a C2 test may still struggle, at least initially, with language use in the types of social contexts associated with non-autonomous language. In any case, it is clear that also under a cognitive processing model of language proficiency, academic language use requires efficient high-level cognitive processing and hence a high level of GLP.

## 3 Academic literacy

The concept of AL is used to describe language skills in higher education contexts. Universities in the UK and US describe AL as a set of study skills that support successful study at undergraduate level and beyond. However, there is variation in how AL is conceptualised. It has been presented as a pre-requisite for starting university study (e.g. ICAS, 2002), whereas others consider it as something primarily developed in early undergraduate study. Bartholomae (1986, p. 4) pointed out that the language demands of university study can be different to those in other contexts:

> [The student] has to learn to speak our language, to speak as we do, to try on the peculiar ways of knowing, selecting, evaluating, reporting, concluding and arguing that define the discourse of our community.

Few academics would dispute Bartholomae's conceptualisation of academic language as something that must be learnt. One key distinction between AL and GLP is that assessments of GLP often conceptualise the native speaker as the highest standard of performance being sought or at least a model of expertise which is typical of native speakers, whereas AL considers the native speaker as a starting point for development. The idea of a native speaker has been discussed extensively in language assessment, but this has not been necessary when considering AL; when it comes to academic language, none of us are native speakers.

### 3.1 Models of academic literacy

There is not universal agreement on what is encompassed by AL. Lea & Street (2006) identified three models of AL: the study skills model, the academic socialisation model and their own academic literacies model (Lea & Street, 1998). The academic literacies approach emphasises social and contextual factors related to writing, and challenges the idea of literacy as a set of generic and transferable cognitive skills (Lea, 2004). Therefore, understandings of AL can be placed on a continuum. At one end are skills-based approaches that conceptualise AL as a set of skills; at the other is the academic literacies approach, which emphasises social and contextual factors related to writing.

To illustrate this continuum, we present three definitions of AL provided by educators. First, the Intersegmental Committee of the Academic Senates (ICAS) published a statement of AL competencies expected of students entering California's public colleges and universities. This defined AL as:

> All the elements of academic literacy—reading, writing, listening, speaking, critical thinking, use of technology, and habits of mind that foster academic success.
>
> (ICAS, 2002, p. 2)

The ICAS definition first lists the four skills typically tested in language assessment contexts. Critical thinking and habits of mind are also mentioned, but the definition is similar to conceptualisations of high-level language proficiency more generally. Another example definition comes from Neeley's (2005, p.7) textbook on AL, which aims to support undergraduate students:

> Academic literacy — proficiency in reading and writing about academic subjects, with the goal of contributing to the ongoing conversations of an academic field.

Neeley also mentions language skills, but focuses on reading and writing, rather than all four skills. The conceptualisation of AL presented by Neeley also adds that development of these skills is linked to a specific goal, extending AL to include socially mediated scholarly activity.

The final example comes from Oxford Brookes' (2014, p.6) Strategy for Enhancing the Student Experience; where AL is listed as a graduate attribute, and defined as:

> Disciplinary and professional knowledge and skills, understanding the epistemology and 'landscape' of the discipline, and what it means to think and behave as a member of that disciplinary and/or professional community of practice.

The Oxford Brookes definition does not explicitly refer to language skills, incorporating this within the concept of disciplinary and professional skills (Sharpe, Benfield, Corrywright & Green, 2014). Compared to the two other examples, this understanding of AL has greater emphasis on disciplines, communities and shared epistemology.

On the continuum, the ICAS (2002) definition represents a skills-based approach, and the Oxford Brookes (2014) one is more compatible with the academic literacies model (Lea & Street, 1998). Although the examples are all from higher education, they are targeted at different stages. This demonstrates how AL can be conceptualised as changing through progressive levels of study, where AL becomes less focused on generic skills, and greater emphasis is placed on nuanced aspects of academic disciplines. AL at later stages of study also focuses on the traditional mediums of communicating in academic communities, resulting in an emphasis on writing.

Whilst the academic literacies model is the dominant approach to higher education writing instruction in the UK (Wingate & Tribble, 2012), it is not suitable for informing assessment practices, because it is critical in emphasis and strongly anti-normative (Lillis, 2003). The rejection of normative academic writing makes it difficult to apply to assessment because there is no normative standard to aim for. Therefore, we focus on AL as a set of skills for the rest of the present paper. However, we acknowledge the limitations of this approach and do not express a preference for the skills-based model of AL in other contexts.

By conceptualising AL as a set of general skills, we can view the construct as having substantial overlaps with GLP. Indeed the ICAS (2002) definition of AL includes the four language skills, and specifically refers to the components that are important for academic study. Even at this level, there are additional elements of AL that do not fall within the remit of GLP. For example, critical thinking and use of technology are related to GLP, but are not considered to be part of the construct itself. Therefore, there are areas of overlap and difference, with greater divergence at more advanced levels.

In the rest of this paper, these overlapping constructs are explored in greater detail to identify areas of further investigation. In particular, we focus on the intersection between these concepts to highlight specific questions about how AL and higher levels of GLP are intertwined.

## 4 Overlap and divergence – some key questions

### 4.1 To what extent is a certain level of general language proficiency a pre-requisite for acquiring academic literacy?

A certain level of GLP is needed before one can engage with the material typically encountered during university study. Lectures include low-frequency words on specialised topics, and academic articles discuss abstract, theoretical concepts. These features mean that a certain level of proficiency is needed for listening and reading study material. For productive language skills, undergraduates are expected to use specialist vocabulary and submit written assignments, which tend to be extended pieces that require structuring and organisation. Group discussions also require students to speak and interact about complex topics.

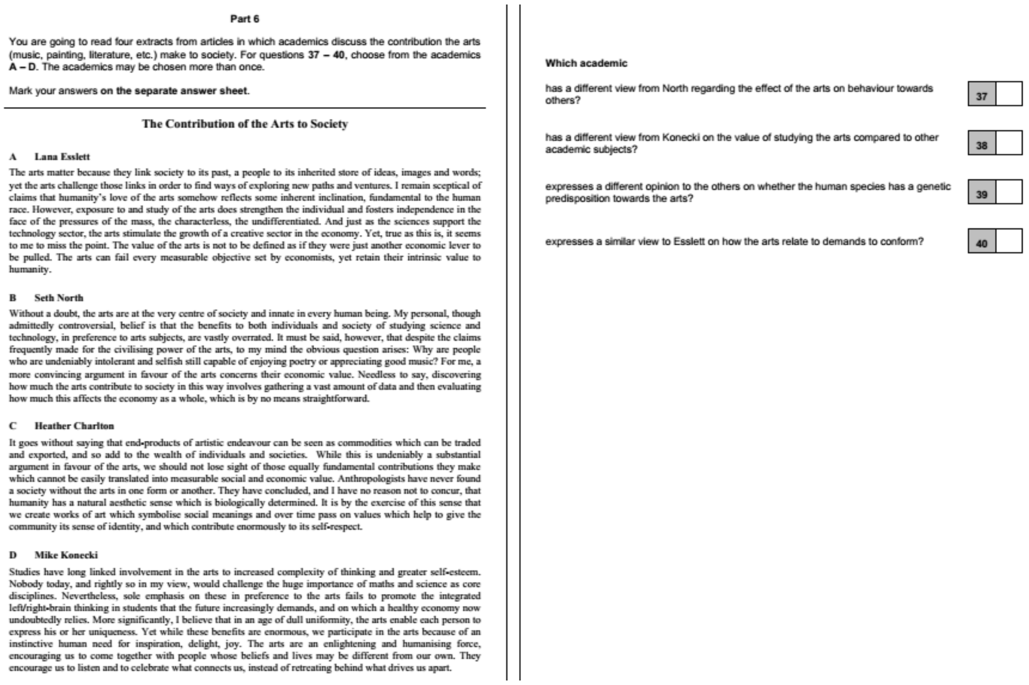
Therefore, it is reasonable to conclude that a certain level of language ability is required for starting to develop AL. However, this assumption that has not been investigated as fully as one might expect. The notion that a threshold level of language proficiency is needed for university study is reflected in GLP admissions criteria for international students. To support decisions about the thresholds that universities set, it would be useful to investigate and identify the language proficiency skills necessary for working with academic material, and importantly, to start developing AL skills. Systematically reviewing CEFR can-do statements could identify appropriate minimum skill requirements.

The development of AL skills in native speakers of English should also be investigated, and compared to the trajectories of those at a range of GLP levels. Another issue arises from the importance placed on English internationally. Non-English speaking countries are encouraging researchers to publish in English and increase their international reach. There is little known about individuals who have developed AL skills in their native language, but have limited proficiency in English. Therefore, the impact of non-English AL on development of English AL is important to consider.

### 4.2 To what extent is high-level language proficiency as conceptualised by the CEFR intrinsically academic?

The CEFR can-do statements describing C1 and C2 show that conceptualisations of GLP at higher levels have an academic component built into them. For example, C2 descriptors for overall reading comprehension (Council of Europe, 2001, p. 69) refer to the handling of ‘abstract, structurally complex’ language and ‘long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning’, while the C1/C2 descriptors for reading for information and argument (Council of Europe, 2001, p. 70) are even more explicit, stating that C-level users can ‘understand in detail a wide range of lengthy, complex texts likely to be encountered in social, professional or academic life, identifying finer points of detail including attitudes and implied as well as stated opinions'. Similarly, the C2 descriptors for writing reports and essays (Council of Europe, 2001, p. 62) state that a C2 user can 'produce clear, smoothly flowing, complex reports, articles or essays which present a case, or give critical appreciation of proposals or literary works' and can 'provide an appropriate and effective logical structure which helps the reader to find significant points'.

The tasks used to assess high-level language proficiency also demonstrate that the CEFR conceptualises these levels as intrinsically academic. For example, Figure 3 shows a reading task from *Cambridge English: Advanced* (C1).



**Part 6**

You are going to read four extracts from articles in which academics discuss the contribution the arts (music, painting, literature, etc.) make to society. For questions **37 – 40**, choose from the academics **A – D**. The academics may be chosen more than once.

Mark your answers **on the separate answer sheet.**

**The Contribution of the Arts to Society**

**A Lana Esslett**

The arts matter because they link society to its past, a people to its inherited store of ideas, images and words; yet the arts challenge those links in order to find ways of exploring new paths and ventures. I remain sceptical of claims that humanity's love of the arts somehow reflects some inherent inclination, fundamental to the human race. However, exposure to and study of the arts does strengthen the individual and fosters independence in the face of the pressures of the mass, the characterless, the undifferentiated. And just as the sciences support the technology sector, the arts stimulate the growth of a creative sector in the economy. Yet, true as this is, it seems to me to miss the point. The value of the arts is not to be defined as if they were just another economic lever to be pulled. The arts can fail every measurable objective set by economists, yet retain their intrinsic value to humanity.

**B Seth North**

Without a doubt, the arts are at the very centre of society and innate in every human being. My personal, though admittedly controversial, belief is that the benefits to both individuals and society of studying science and technology, in preference to arts subjects, are vastly overrated. It must be said, however, that despite the claims frequently made for the civilising power of the arts, to my mind the obvious question arises: Why are people who are undeniably intolerant and selfish still capable of enjoying poetry or appreciating good music? For me, a more convincing argument in favour of the arts concerns their economic value. Needless to say, discovering how much the arts contribute to society in this way involves gathering a vast amount of data and then evaluating how much this affects the economy as a whole, which is by no means straightforward.

**C Heather Charlton**

It goes without saying that end-products of artistic endeavour can be seen as commodities which can be traded and exported, and so add to the wealth of individuals and societies. While this is undeniably a substantial argument in favour of the arts, we should not lose sight of those equally fundamental contributions they make which cannot be easily translated into measurable social and economic value. Anthropologists have never found a society without the arts in one form or another. They have concluded, and I have no reason not to concur, that humanity has a natural aesthetic sense which is biologically determined. It is by the exercise of this sense that we create works of art which symbolise social meanings and over time pass on values which help to give the community its sense of identity, and which contribute enormously to its self-respect.

**D Mike Konecki**

Studies have long linked involvement in the arts to increased complexity of thinking and greater self-esteem. Nobody today, and rightly so in my view, would challenge the huge importance of maths and science as core disciplines. Nevertheless, sole emphasis on these in preference to the arts fails to promote the integrated left/right-brain thinking in students that the future increasingly demands, and on which a healthy economy now undoubtedly relies. More significantly, I believe that in an age of dull uniformity, the arts enable each person to express his or her uniqueness. Yet while these benefits are enormous, we participate in the arts because of an instinctive human need for inspiration, delight, joy. The arts are an enlightening and humanising force, encouraging us to come together with people whose beliefs and lives may be different from our own. They encourage us to listen and to celebrate what connects us, instead of retreating behind what drives us apart.

**Which academic**

has a different view from North regarding the effect of the arts on behaviour towards others? 37 □

has a different view from Konecki on the value of studying the arts compared to other academic subjects? 38 □

expresses a different opinion to the others on whether the human species has a genetic predisposition towards the arts? 39 □

expresses a similar view to Esslett on how the arts relate to demands to conform? 40 □

**Figure 3.** An example reading task from a *Cambridge English: Advanced paper*.

In this task, the test taker must read extracts from articles written by academics, and then compare the views that these authors express on various topics. This kind of language skill is used in university study contexts to establish scholarly arguments made in a range of subjects. Furthermore, there are not many contexts outside of academic study or scholarly practice where this would be employed regularly.

The task demonstrates how the CEFR's conceptualisation of high-level language ability is situated within academic contexts, and how they are preferred over other uses of language. There are complex language skills that are arguably more difficult to demonstrate. Writing lyrics for a song, telling jokes or raising queries after reading a tenancy agreement are all tasks that potentially demonstrate high levels of language proficiency. However, they are not situations commonly used in language tests. Given the influence of higher education on language learning, and vice versa, the focus on academic language in the CEFR is understandable and, from many perspectives, a suitable approach to adopt. Most individuals seeking to learn language with a formal framework are doing so for educational or professional purposes. Indeed, the practical use of a test assessing joke-telling ability would be limited.

Although it is clear that higher levels of GLP have academic elements, it important to acknowledge that the emphasis can differ. For example, C1 and C2 speaking tasks tend to focus on more general topics, and successfully elicit high-level language using interaction. It would be misleading to claim that language proficiency at the C1 and C2 levels are entirely academic. Therefore, it would be useful to review tasks and CEFR can-do statements alongside definitions of AL, and the skills used in undergraduate study. Another area to investigate is how native speakers without experience of academic study perform on high-level language tests. It is possible that C2 is difficult to achieve without prior experience of further education.

### 4.3 In what ways can academic literacy be considered an extension of language proficiency as conceptualised by the CEFR?

The examples shown so far characterise AL as an extension of GLP. AL, at least using the skills-based model, could be seen as another level on the CEFR. However, this would be an oversimplification of the relationship between these constructs.

Table 2 is a set of example essay questions from an undergraduate module previously taught by one of the authors. These illustrate important differences between AL and GLP. Firstly, discipline-specific knowledge is needed to demonstrate AL in a university setting. It is difficult to attempt these questions without an understanding of intelligence theories, the five factor model of personality, or some familiarity with Eysenck's work. Therefore, academic assignments require shared understandings of a subject area, because the tasks aim to assess domain-specific knowledge and understanding, rather than AL.

| Example undergraduate writing assignments |
|---|
| Critically appraise the contribution of psychological testing to the development of theories about intelligence. |
| To what extent can the five factor model be considered a universal model of personality? |
| Critically evaluate H.J. Eysenck's concept of extraversion. |

**Table 2.** Example assignment questions from an undergraduate psychology module

Another observation is that academic assignments tend to only require a written submission. Although students are expected to read relevant materials, reading is not directly assessed. Listening to lectures and talking about the topics facilitate the development of subject knowledge, but these language skills are not explicitly included in the assessment.

Undergraduate students find academic writing particularly difficult to develop, because it involves complex skills that are assessed using criteria they are unlikely to have encountered before (Elander, Harrington, Norton, Robinson & Reddy, 2006). Therefore, a focus on writing over the other three language skills is understandable.

Elander, Harrington, Norton, Robinson, Reddy & Stevens (2004) identified a core set of four assessment criteria that are used to assess academic writing across disciplines. Two relate to sub-skills typically included in language tests for speakers of other languages – 'use of language/writing style' and 'structuring'. The other two, 'critical thinking/critical evaluation' and 'developing argument', represent skills that are specifically valued in higher education settings.

The importance of critical thinking and argument is reflected in definitions of AL and researchers in the US have also identified core criteria that include these domains (e.g. Dryer, 2013). Furthermore, higher education study introduces concepts specific to AL contexts, such as plagiarism and authorial identity (Pittam, Elander, Lusher, Fox & Payne, 2009). Recently, researchers have investigated these aspects of academic study (e.g. Cheung, Elander, Stupple & Flay, 2016; Stupple, Maratos, Elander, Hunt, Cheung & Aubeeluck, 2017; Zhao, 2013); however, cognitive processing models that include these skills have not been developed, presenting an area for further research.

We have argued that AL is not merely an extension of GLP, because it includes skills outside of the language domain that are socially mediated, discipline-specific and have not been included in cognitive processing models of GLP. As AL skills enable students to demonstrate subject-specific knowledge in written assignments, the construct is typically assessed in contexts where it is necessary, but not sufficient, for success.

### 4.4 Can language testers avoid conflating the concepts of EAP and general language proficiency?

Due to the CEFR's emphasis on academic contexts at high-levels of language proficiency, language tests targeting language proficiency at these levels will unavoidably include tasks typically associated with academic study. However, this is not problematic. In fact, including some aspects of AL is necessary when assessing language at higher levels, because excluding them would compromise construct coverage.

Overlap of the constructs in an assessment may be suitable, but an important issue to consider is the degree of overlap that is suitable. Crucially, establishing the components of AL to include in language tests will inform assessment design. Furthermore, systematic investigation of this area can identify the limitations of language tests, and more explicitly define the expectations that university stakeholders should have of students entering undergraduate study.

## 5 Implications for testing

Examples from a range of sources have been used to compare high level language proficiency and AL, identifying similarities and differences (see Figure 4). However, this comparison is not proposed as a comprehensive one; instead, they are intended to prompt further discussion. Similarly, the questions we have posed and our suggestions for investigating them are not authoritative. Instead, we invite a critical approach to evaluating the assumptions that they rest upon.

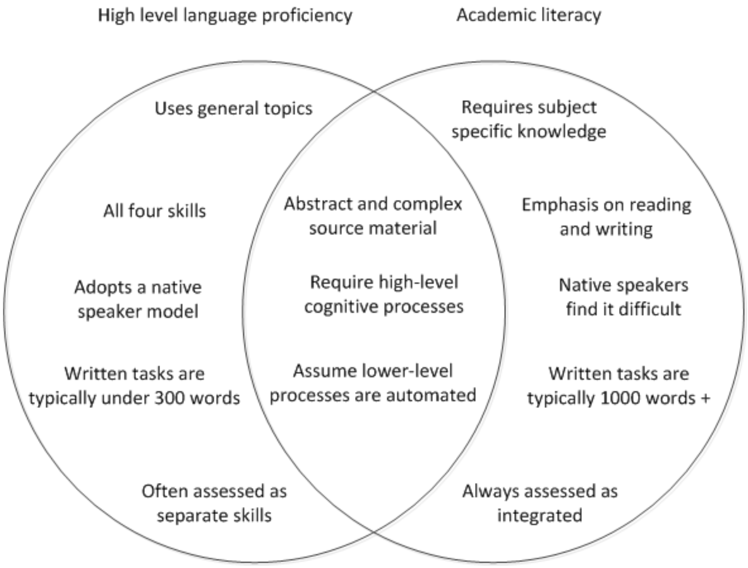

High level language proficiency
Academic literacy
Uses general topics
All four skills
Adopts a native speaker model
Written tasks are typically under 300 words
Often assessed as separate skills
Abstract and complex source material
Require high-level cognitive processes
Assume lower-level processes are automated
Requires subject specific knowledge
Emphasis on reading and writing
Native speakers find it difficult
Written tasks are typically 1000 words +
Always assessed as integrated

**Figure 4.** A comparison of high-level language proficiency and academic literacy

Investigating the relationship between AL and GLP may also help develop language tests that are more sensitive to the needs of university study. Murray and Nallaya (2016) highlighted that some students face language problems at university, despite meeting GLP entry criteria. They advocate assuming that all students, including those with English as a first language, need to develop familiarity with academic language. Demonstrating that native English speakers need support to develop AL would empirically contribute to the arguments for these interventions. Assessments designed specifically for evaluating AL could be administered to native English speakers, to support these kinds of approaches.

Many approaches to developing students' academic writing draw on Lea and Street's (2006) academic literacies model, which rejects skills-focused approaches to AL (Wingate, 2012). However, conceptualising *some* aspects of AL as skills to assess has benefits; for example, assessment tools can standardise feedback for specified areas so that instructors devote time to commenting on more nuanced and subject-specific aspects of AL. Importantly, this approach recognises that some aspects of AL are unsuitable for assessment as a set of generic skills. The resource demands of providing feedback on writing tasks are a potential barrier to wider use of embedded writing instruction (Wingate, Andon & Cogo, 2011); therefore, improving their efficiency with limited testing might make them more feasible with larger courses.

The present paper does not purport to offer definitive solutions to the issues outlined in necessarily sketchy detail – rather, it intends to present some issues and strands of thinking we have been grappling with while trying to accommodate and disentangle the two concepts of GLP and AL, while providing some analysis of key issues, heuristics for approaching the topic and areas for further research. It is clear that there is both overlap and divergence between the two concepts, an observation which must have implications for language testers; the task is to identify those similarities and differences in detail, determine the consequences and use this information to guide test design. This paper is an attempt to contribute to this process, in the hope that others may succeed in unravelling some of the more complex issues we have yet to fully unravel ourselves.

## References

Bartholomae, D. (1986). Inventing the University. *Journal of Basic Writing, 5*(1), 4–23.

Bernstein, B. (1964) Elaborated and restricted codes: their social origins and some consequences. *American Anthropologist, 66*(6), 55–69.

Cheung, K. Y. F., Elander, J., Stupple, E. J. N., & Flay, M. (2016). Academics' understandings of the authorial academic writer: A qualitative analysis of authorial identity. *Studies in Higher Education*, 1-16. DOI: 10.1080/03075079.2016.1264382

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Dryer, D. B. (2013). Scaling Writing Ability: A Corpus-Driven Inquiry. *Written Communication, 30*(1), 3–35.

Elander, J., Harrington, K., Norton, L., Robinson, H., & Reddy, P. (2006). Complex skills and academic writing: a review of evidence about the types of learning required to meet core assessment criteria. *Assessment & Evaluation in Higher Education, 31*(1), 71–90.

Elander, J., Harrington, K., Norton, L., Robinson, H., Reddy, P., & Stevens, D. (2004). Core assessment criteria for student writing and their implications for supporting student learning, in C. Rust (Ed.), *Improving Student Learning 11: Theory, Research and Scholarship* (pp. 200–212). Oxford: The Oxford Centre for Staff and Learning Development.

Elliott, M. (2011). *FCE and CAE Construct Validation Study (Part 1)*. Cambridge English Language Assessment internal report 1356.

Field, J. (2011). Cognitive validity, in L. Taylor (Ed.), *Examining Speaking: Research and practice in assessing second language speaking* (pp. 65–111). Cambridge: UCLES/Cambridge University Press.

Field, J. (2013). Cognitive validity, in A. Geranpayeh, A & L. Taylor, L (Eds.), *Examining Listening: Research and practice in assessing second language listening* (77–151). Cambridge: UCLES/Cambridge University Press.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*(4), 663–667.

ICAS. (2002). *Academic Literacy: A Statement of Competencies Expected of Students Entering California's Public Colleges and Universities*. California: ICAS. Retrieved from: http://icas-ca.org/academic-literacy

Jessen, A. & Elander, J. (2009). Development and evaluation of an intervention to improve Further Education students' understanding of Higher Education assessment criteria: three studies. *Journal of Further and Higher Education*, *33*, 359–380.

Kay, P. (1977). Language evolution and speech style. in B. G. Blount, & M. Sanches (Eds), *Sociolinguistic Dimensions of Language Change* (pp. 21–33). New York: Academic Press.

Khalifa, H. & Weir, C. J. (2009). *Examining Reading: Research and practice in assessing second language reading*. Cambridge: UCLES/Cambridge University Press.

Klein, W. & Perdue, C. (1997). The basic variety (or: Couldn't natural languages be much simpler?). *Second Language Research, 13*(4), 301–347.

Lea, M. R. (2004). Academic literacies: a pedagogy for course design. *Studies in Higher Education, 29*(6), 739–756.

Lea, M. R. & Street, B. (1998). Student Writing in Higher Education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157-172.

Lea, M. R. & Street, B. (2006). The "Academic Literacies" model: Theory and applications. *Theory into Practice, 45*(4), 368–377.

Lillis, T. (2003). Student writing as ‘Academic Literacies’: Drawing on Bakhtin to move from critique to design. *Language and Education, 17*(3), 192–207.

Murray, N. & Nallaya, S. (2016). Embedding academic literacies in university programme curricula: a case study. *Studies in Higher Education, 41*(7), 1,296–1,312. DOI: 10.1080/03075079.2014.981150.

Neeley, S.D. (2005). *Academic Literacy*. London: Pearson.

Oxford Brookes University. (2014). *Strategy for Enhancing the Student Experience 2 (2015–2020)*. Retrieved from: http://www.brookes.ac.uk/ocsld/your-development/teaching-and-learning/graduate-attributes/

Pittam, G., Elander, J., Lusher, J., Fox, P., & Payne, N. (2009). Student beliefs and attitudes about authorial identity in academic writing. *Studies in Higher Education, 34* (2), 153–170.

Sharpe, R., Benfield, G., Corrywright, D., & Green, L. (2014). *Evaluation of the Brookes Graduate Attributes: Year 1 Final Report*. Retrieved from: https://wiki.brookes.ac.uk/download/attachments/120946694/GAsEvaluationFinalv4.3.pdf

Shaw, S. D. & Weir, C. J. (2007). *Examining Writing: Research and practice in assessing second language writing*. Cambridge: UCLES/Cambridge University Press.

Stupple, E. J. N., Maratos, F. A., Elander, J., Hunt, T. E., Cheung, K. Y. F., & Aubeeluck, A. V. (2017) Development of the Critical Thinking Toolkit (CriTT): A measure of student attitudes and beliefs about critical thinking, *Thinking Skills and Creativity*, *23*, 91–100.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

Wingate, U. (2012). Using Academic Literacies and genre-based models for academic writing instruction: A ‘literacy’ journey, *Journal of English for Academic Purposes, 11*, 26-37.

Wingate, U. & Tribble, C. (2012). The best of both worlds? Towards an English for Academic Purposes/Academic Literacies writing pedagogy. *Studies in Higher Education.* 37(4), 481–495. DOI: 10.1080/03075079.2010.525630

Wingate, U., Andon, N., & Cogo, A. (2011). Embedding academic writing instruction into subject teaching: A case study. *Active Learning in Higher Education, 12*(1), 69–81.

Zhao, C. F. (2012). Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. *Language Testing, 30*(2), 201–230.

## Language learning, teaching and assessment…

## … in a multilingual world

# Mediation and Exploiting One's Plurilingual Repertoire: Exploring Classroom Potential with Proposed New CEFR Descriptors

**Brian North**, Eurocentres Foundation, Switzerland
**Enrica Piccardo**, OISE – University of Toronto, Canada

**Abstract:** This paper reports on a 2014–2017 Council of Europe project to update the CEFR's 2001 illustrative descriptor scales with descriptors for areas that were not covered in the original set, namely mediation, online interaction, reactions to literature, and plurilingual and pluricultural competences. The approach taken to mediation is far broader than in some interpretations. In the set of descriptors, mediating concepts and mediating communication are covered in addition to cross-linguistic mediation of a text. The development followed the same three phase process as in the original CEFR descriptor research in the 1993–6 Swiss research project (intuitive authoring and editing, qualitative validation in workshops, quantitative validation through Rasch model scaling), but on a larger scale. Some 150 institutions and 1,300 people took part in the three validation phases that took place between February 2015 and February 2016, with a formal consultation phase from October 2016 till February 2017.

## 1 Introduction

In its move beyond the four skills, the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) pioneered the introduction of mediation as a fourth mode of communication alongside reception, production and interaction. The CEFR gives mediation a key role in its action-oriented approach and shows awareness of the way mediation spans the linguistic, cultural and social dimensions (Piccardo, 2012). The CEFR stresses the social, collaborative vision of language by seeing the user/learner as a social agent and – in mediation – as an intermediary between different interlocutors who are unable, for whatever reason, to communicate directly (Council of Europe 2001, p. 14). A second notion introduced in the CEFR is plurilingualism, seen as an uneven, dynamic competence in which capacities in each language may be very different. Partial competences in different languages are presented as being of great value, as a stepping-stone to further development. Taken together, the pioneering notions of mediation and plurilingualism offer a paradigm shift in language education.

## 2 Mediation

An increasing awareness of the complex nature of the process of second language teaching and learning has led to a growing interest in recent years in the notions of mediation and plurlingualism. Mediation involves the use of language in creating the space and conditions for communication and/or learning, in constructing and co-constructing new meaning, and/or in facilitating understanding by simplifying, elaborating, illustrating or otherwise adapting the original. In fact, though, mediation has been seen from many different perspectives, being described as a "nomadic notion" (Lenoir, 1996), a term used in different senses in different contexts. The use of mediation in relation to diplomacy, conflict resolution and commercial transaction has crossed the ages from the classical to the contemporary world while expanding to include a wide range of professional arbitration, counselling and guidance activities. Our deeper reflection on the nature of mediation, though, is rooted in philosophy, namely in German idealism and dialectical materialism. For Hegel, thought was a mediation process, an abstract operation through which knowledge was acquired, a view to which Marx and Engels added a social dimension in which mediation was a form of relation between opposing domains and forces in the society. This trailblazing understanding of the twofold nature of mediation informed reflection in a broad range of disciplines. In particular, the work of Vygotsky (1978) enabled the crucial transition to psychology and education by explaining how social interaction plays a fundamental role in the development of cognition. Multifaceted in its nature, mediation always implies a process that can be either social in nature or situate itself at the level of the individual. In the former case, it focuses on scaffolding (Wood, Bruner & Ross, 1976), enhancing communication and reciprocal comprehension, or on bridging gaps and resolving tensions. In the latter, it concerns psycho-cognitive development. However, an awareness of the interdependence of these two dimensions – individual and collective, cognitive and social – has been crucial in understanding mediation. The sociocultural view of learning stresses this interdependence (Lantolf, 2000), seeing mediation at the core of knowledge (co)construction and language at the core of mediation. Indeed, language itself can transform into a process: languaging, “a dynamic, never-ending process of using language to make meaning” (Swain, 2006, p. 96).

## 3 Mediation in the CEFR

The CEFR does not greatly develop the notion of mediation, which was introduced to replace the fourth mode of communication “processing”, intended to cover communicative language activities involving integrated skills and summarising that were not covered by the trio reception, interaction and production that North proposed (1992; 1994). Essentially the interpretation of mediation in the CEFR text arose by the addition of a cross-linguistic focus to that processing. However, the presentation of mediation in the CEFR goes considerably beyond information transfer and (professional) interpretation/translation, as can be seen from the following two extracts. The concept is introduced at the beginning of the CEFR as follows:

> In both the receptive and productive modes, the written and/or oral activities of *mediation* make communication possible between persons who are unable, **for whatever reason** to communicate with each other directly. Translation or interpretation, a paraphrase, summary or record, provides for a third party a (re)formulation of a source text to which this third party does not have direct access. Mediation language activities, (re)processing an existing text, occupy an important place in the normal linguistic functioning of our societies.
>
> (CEFR Section 2.1.3: English p. 14, French p. 18: emphasis added)

This is elaborated further when discussing mediation in more detail:

> In mediating activities, the language user is not concerned to express his/her own meanings, but simply to act as an intermediary between interlocutors who are unable to understand each other directly – normally (but not exclusively) speakers of different languages. Examples of mediating activities include spoken interpretation and written translation as well as summarising and paraphrasing texts in the same language, when the language of the original text is not understandable to the intended recipient.
>
> (CEFR Section 4.4.4: English p. 87, French p. 71: emphasis added)

In the CEFR as a whole, at least four different mediation situations, which are in practice often combined, are mentioned. In these activities, the user/learner:

- receives a text and produces a related text to be received by another person who has no access to the first text;
- acts as an intermediary in a face-to-face interaction between two interlocutors who do not understand one another, possibly because they do not share the same language or code;
- interprets a cultural phenomenon in relation to another culture;
- participates in a conversation or discussion that involves several languages, exploiting his/her plurilingual and pluricultural repertoires.

In addition, the CEFR emphasises the two key notions of co-construction of meaning in interaction and constant movement between the individual and social level in language learning, mainly through its vision of the user/learner as a social agent (Piccardo, 2012). The CEFR stresses how the external context must always be interpreted by the user/learner and also reminds us that there is a form of interior mediation that takes place at the level of the individual. The social agent and his/her interlocutor share the same situational context but may well maintain different perceptions and interpretations. The gap between these may be so great as to require some form of mediation, perhaps even by a third person. This view is in fact very compatible with several recent approaches to second language learning, especially approaches informed by sociocultural and socio-constructivist theories (Lantolf, 2000; Schneuwly, 2008), in which mediation is a key concept.

To summarise, the CEFR touches upon several different aspects of mediation: textual and cross-linguistic mediation (Backus et al., 2013; Stathopoulou, 2015), social and cultural mediation (Zarate, 2003; Zarate, Gohard-Radenkovic, & Lussier, 2004) and conceptual mediation (Dawson, 2014), both scaffolded (Walqui, 2006; Zwiers, 2008) and through collaborative learning (Barnes & Todd, 1977; Mercer & Dawes, 2008; Webb, 2009).

## 4 Descriptor scales for mediation and plurilingualism

The context to the development of CEFR descriptors for mediation and plurilingual/pluricultural competence is an initiative of the Council of Europe to commission an update of the CEFR's illustrative descriptors. The project had several facets, and only the second one is reported on in this paper:

(1) improve the coverage at A1 and the C levels, enrich the description of listening and reading by profiting from the various projects that have validated and calibrated descriptors to the CEFR levels following an approach similar to that in the original CEFR descriptor research (North, 2000; North and Schneider, 1998), and replace the scale for phonological control (Piccardo & North, 2017).

(2) develop, validate and calibrate descriptors for new areas, particularly online interaction, mediation, plurilingual/pluricultural competence, and reactions to literature (North & Piccardo, 2016).

(3) incorporate descriptors for sign languages, mainly based on a Swiss National Science Research Council project (Keller, Meili, Bürgin, & Ni, 2017); these will be added in January 2018.

(4) provide a collation of descriptors for young learners, related to the extended version of the illustrative descriptors (Szabo & Goodier, 2017).

The view taken of mediation is a relatively broad one, encompassing the aspects listed in the previous section, grouped under mediating a text, mediating concepts and mediating communication. During the development, an authoring group consisting of Brian North, Tim Goodier, Enrica Piccardo and Maria Stathopoulou was accompanied by a 'sounding board' giving interactive feedback, plus a group of consultants. The categories for these new descriptor scales developed in the project are as follows. Italics indicate a title for a group of scales, not a scale.

*Online interaction*

Online conversation and discussion

Goal-oriented online transactions and collaboration

*Mediation*

Overall mediation

*Mediating a text*

Relaying specific information in speech

Relaying specific information in writing

Explaining data (e.g. in graphs, diagrams, charts etc.) in speech

Explaining data (e.g. in graphs, diagrams, charts etc.) in writing

Processing text in speech

Processing text in writing

Translating a written text in speech

Translating a written text in writing

Note-taking (lectures, seminars, meetings, etc.)

Expressing a personal response to creative texts (including literature)

Analysis and criticism of creative texts (including literature)

*Mediating concepts*

*Collaborative work within a group*

Facilitating collaborative interaction with peers

Collaborating to construct meaning

*Leading group work*

Managing interaction

Encouraging conceptual talk

*Mediating communication*

Facilitating pluricultural space

Acting as intermediary in informal situations (with friends and colleagues)

Facilitating communication in delicate situations and disagreements

*Mediation strategies*

*Strategies to explain a new concept*

Linking to previous knowledge

Adapting language

Breaking down complicated information

*Strategies to simplify a text*

Amplifying a dense text

Streamlining a text

*Plurilingual and pluricultural competence*

Building on pluricultural repertoire

Plurilingual comprehension

Building on plurilingual repertoire

## 5 Validation

The project emulated and further extended the methodologies employed in the original CEFR descriptor research (North, 2000; North & Schneider, 1998), following a similar mixed method (Cresswell, 2003), qualitative and quantitative developmental research (Richey & Klein, 2005) design. Three phases of validation were carried out between February and November 2015, with a further phase of validation for plurilingual/pluricultural in January–February 2016. In each phase, a data collection matrix was used to ensure that each of the descriptors for each draft scale was evaluated by between 150 and 250 persons, with a representative distribution across countries and educational contexts.

### 5.1 Phase 1: Qualitative

137 institutes and circa 990 respondents took part in a series of workshops at their institutions. The task was to assign descriptors to the scale to which they belonged, to evaluate them for clarity, pedagogical usefulness and relation to real world language use, and to propose improvements to the formulation. Respondents often radically shortened descriptors, confirming North’s (2000, p. 345) finding that teachers prefer descriptors of up to 20 words.

### 5.2 Phase 2: Qualitative and quantitative

189 institutions from 45 countries and 1294 persons took part in the second series of workshops. Respondents assigned descriptors to CEFR levels by answering the question: At what CEFR level do you think a person can do what is defined in the descriptor? Each participant marked their decisions first on paper and then, after discussion with their partner, reflection and review, entered a final judgement into a SurveyMonkey. As well as considering the percentages who selected the intended level, and the spread across levels for individual descriptors, a Rasch analysis (Linacre, 2015) was also carried out, comparing three different ways to “anchor” the descriptors to the scale underlying the CEFR levels (North, 2000).

### 5.3 Phase 3: Quantitative

An online survey was then used to replicate the calibration task from the original CEFR descriptor research. Respondents were asked to think how a person that they knew very well (themselves or someone else) would perform in relation to each descriptor, answering the question: Could you, or the person concerned, do what is described in the descriptor? using the same 0–4 rating scale that had been used in the original CEFR descriptor research (North, 2000; North & Schneider, 1998).

### 5.4 Further validation for plurilingual/pluricultural

Finally, an extra survey was carried out in February 2016 for plurilingualism. The opportunity was taken to also include descriptors for reception strategies and plurilingual comprehension, and to add more descriptors for pluricultural competence, particularly at lower levels. The survey was carried out in two parallel versions. 267 volunteers from among the project participants completed one form, whilst 62 experts in plurilingual education completed the other. The results proved identical from both groups and the calibrations to level were also extremely compatible with the existing CEFR scale for sociolinguistic appropriateness.

## 6 Conclusion

Following the development and validation, a process of systematic consultation took place between July 2016 and February 2017. After an expert meeting and a pre-consultation of experts, a formal consultation exercise was undertaken with member states, institutions and individuals. Respondents were asked a series of questions including how helpful the various new scales were. Responses from institutions and the over 500 individuals were overwhelmingly positive and gave very useful suggestions for further editing of some of the descriptors. The most ‘popular’ categories were plurilingual/pluricultural (member states) mediating a text, collaborative work within a group and online interaction. There was a noticeable difference of opinion between individuals and institutions on just two descriptor scales: Goal-oriented online transactions and collaboration and Building on plurilingual repertoire. Whilst 96% of the institutions found these two scales helpful or very helpful, only 81% of individuals did so.

Piloting has been taking place since January 2017, with some 55 pilots completed at the time of writing. The most popular areas were collaborative work within a group, mediating a text and plurilingual/pluricultural competence. The vast majority of the pilots selected descriptors from relevant scales in order to inform the design of communicative tasks in the classroom, and then used the descriptors to observe the language use of the learners. Feedback on the descriptors was very positive, with some useful suggestions for small revisions. The product from the project, a CEFR Companion Volume with an extended version of the illustrative descriptors, has been online in a provisional English edition since August 2017. Properly published versions in English, French and German, incorporating sign languages, should be available from very early 2018. Further experimentation with the descriptors, particularly exploration of their relevance to different educational sectors, will be ongoing in the academic year 2017–18.

The Authoring Group hope that the provision of CEFR descriptors for mediating text, mediating concepts, mediating communication and for plurilingual/pluricultural competence will help to broaden the types of tasks carried out in language classrooms and to value all the developing language resources that users/learners bring. The Council of Europe hopes that the Companion Volume, with its extension of the CEFR illustrative descriptors to include areas such as mediation, plurilingual/pluricultural competence and sign languages will contribute to the inclusive right to Quality Education for all, and the promotion of plurilingualism and pluriculturalism.

It is important to note that the Companion Volume, and in particular the descriptors for new areas, represent an enrichment of the original CEFR descriptive apparatus, not a replacement. The additions do not impact on the construct described in the CEFR, or on its Common Reference Levels. Considerable care was taken to ensure that the descriptors for the new areas are calibrated accurately to the original scale underlying the CEFR levels. In addition to the Companion Volume itself, a full report (North & Piccardo, 2016) and background technical reports are available on the Council of Europe's CEFR website.

### References

Backus, A., Gorter, D., Knapp, K., Schjerve-Rindler, R., Swanenberg, J., ten Thije, J. D., & Vetter, E. (2013). Inclusive Multilingualism: Concept, Modes and Implications. *European Journal of Applied Linguistics, 1*(2), 179–215. https://doi.org/10.1515/eujal-2013-0010

Barnes, D., & Todd, F. (1977). *Communication and learning in small groups*. London, UK: Routledge and Kegan Paul.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Cresswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Dawson, C. (2014) Towards a conceptual profile: Rethinking conceptual mediation in the light of recent cognitive and neuroscientific findings. *Research in Science Education, 44*, 389–414. DOI 10.1007/s11165-013-9388-4

Keller, J., Meili, A., Bürgin, P. S., & Ni, D. (2017). Auf dem Weg zum Gemeinsamen Europäischen Referenzrahmen (GER) für Gebärdensprachen: Empirie-basierte Bestimmung von Deskriptoren für Textkompetenz am Beispiel der Deutschschweizer Gebärdensprache (DSGS). *Das Zeichen 105*, 86–97. Retrieved from http://www.idgs.uni-hamburg.de/de/forschung/publikationen/daszeichen.html

Lantolf, J. P. (2000). *Sociocultural theory and second language learning*. Oxford, UK: Oxford University Press.

Lenoir, Y. (1996). Médiation cognitive et médiation didactique. In C. Raisky & M. Caillot (Eds.), *Le didactique au delà des didactiques. Débats autour de concepts fédérateurs* (pp. 223–251). Bruxelles: De Boeck Université.

Linacre, J. M. (2015). *Winsteps: Rasch-model computer program*, Chicago, IL: MESA Press.

Mercer, N., & Dawes, L. (2008). The value of exploratory talk. In N. Mercer, & S. Hodgkinson (Eds.), *Exploring talk in schools* (pp. 55–72). London, UK: Sage.

North, B. (1992). European Language Portfolio: Some options for a working approach to design scales for proficiency. In: *Council of Europe Transparency and coherence in language learning in Europe: Objectives, assessment and certification*. Symposium held in Rüschlikon, 10–16 November 1991 (pp. 158–174). Strasbourg: Council for Cultural Co-operation. Reprinted in Schärer, R., & North, B. (1992). *Towards a common European framework for reporting language competency*. Washington, DC: NFLC Occasional Paper, National Foreign Language Center, April 1992.

North, B. (1994). *Perspectives on Language Proficiency and Aspects of Competence: a reference paper defining categories and levels*. Strasbourg: Council of Europe CC-LANG (94) 20.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.

North, B. & Piccardo, E. (2016): Developing illustrative descriptors of aspects of mediation for the Common European Framework of Reference (CEFR). Research report. Strasbourg, France: Council of Europe, Language Policy Unit.

North, B. & Schneider, G (1998). Scaling descriptors for language proficiency scales, *Language Testing 15* (2), 217–262.

Piccardo, E. (2012). Médiation et apprentissage des langues: Pourquoi est-il temps de réfléchir à cette notion ? In J. Aden & D. Weissmann (Eds.), *ÉLA (Études de Linguistique Appliquée) 167*, Didier édition Klincksieck, 285–297.

Piccardo, E. & North, B. (2017). Developing phonology descriptors for the Common European Framework of Reference (CEFR). In M. O'Brien & J. Levis (Eds.), *Proceedings of the 8th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Calgary, AB, August 2016 (pp. 97–109). Ames, IA: Iowa State University.

Richey, R. C. & Klein, J. D. (2005). Developmental research methods: Creating knowledge from instructional design and development practice. *Journal of Computing in Higher Education 16*(2), 23–38.

Schneuwly, B. (2008). Vygotski, l'école et l'écriture. *Cahiers des Sciences de l'éducation, 118*. Geneva: Université de Genève.

Stathopoulou, M. (2015). *Cross-language mediation in foreign language teaching and testing*. Cleveland: Multilingual Matters.

Swain, M. (2006). Languaging, agency and collaboration in advanced language proficiency. In H. Byrnes (Ed.), *Advanced language learning: The contribution of Halliday and Vygotsky* (pp. 95–108). London, UK & New York, NY: Continuum.

Szabo, T. & Goodier, T. (2017). *Collated representative samples of descriptors of language competences developed for young learners: Resource for educators*. Version 1 developed through Eurocentres consultancy for the Council of Europe. Strasbourg, France: Council of Europe. Retrieved from https://mycloud.coe.int/index.php/s/I9NfLJPAECo0jOr

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Walqui, A. (2006). Scaffolding instruction for English language learners: A conceptual framework. *The International Journal of Bilingual Education and Bilingualism 9*(2), 159–180.

Webb, N. (2009). The teacher's role in promoting collaborative dialogue in the classroom. *British Journal of Educational Psychology 78*(1), 1–28.

Wood, D. J., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychiatry and Psychology 17*(2), 89–100.

Zarate, G. (2003). Identities and plurilingualism: Preconditions for the recognition of intercultural competences. In M. Byram (Ed.), *Intercultural competence* (pp. 84–117). Strasbourg, France: Language Policy Division, DG IV – Directorate of School, Out-of-School and Higher Education, Council of Europe.

Zarate, G., Gohard-Radenkovic, A., & Lussier, D. (2004). *Cultural mediation in interculturalism and multiculturalism: Similarities and differences*. Strasbourg, France: Council of Europe Publishing.

Zwiers, J. (2008). *Building academic language*. San Francisco, CA: Jossey-Bass.

# QualiCEFR: A Quality Assurance Template to Achieve Innovation and Reform in Language Education through CEFR Implementation

**Enrica Piccardo**, OISE-University of Toronto, Canada
**Brian North**, Eurocentre Foundation, Switzerland
**Eleonora Maldina**, OISE-University of Toronto, Canada

**Abstract:** The CEFR has informed teaching, assessment and testing practices worldwide. Yet, its implementation is largely uninformed by Quality Assurance (QA) or impact studies. This article reports on QualiCEFR, a two-year international comparative research study funded by the Social Science and Humanities Research Council of Canada, integrating qualitative and quantitative research methods with a QA approach to inform and improve CEFR implementation. The project consists of two phases: firstly, a comparison between Switzerland and Canada, two multilingual countries with decentralized education systems, the former having been at the forefront of CEFR implementation, the latter in the earlier stages. The focus is on transparency and coherence in language curriculum reform and teacher development, appropriateness of QA procedures, and identification of successes and challenges in the CEFR implementation process. Over 40 interviews with key players have been conducted and thematically analyzed. CEFR-related initiatives, promising practices and implementation outcomes that can be replicated and upscaled are being identified.

## 1 The QualiCEFR Project: reasons and aims

In our knowledge society, foreign languages play a major role in innovation, competitiveness, and productivity, facilitating globalized communication and mobility. However, in many countries, foreign language proficiency is generally modest, despite much time dedicated to language learning in school curricula. In 2001, the Council of Europe published the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) to address this problem. The CEFR is a language policy document intended to define levels of language proficiency in terms of real-world practical ability, stimulate educational reform, and provide coherence between curricula, teaching practices and evaluation.

Byram and Parmenter (2012) have shown that the CEFR has increasingly informed reform of pedagogy and assessment practices worldwide. However, the implementation of the CEFR has not proceeded systematically, but rather has been left to the uncoordinated initiatives of national, regional, and/or local authorities. Very few impact studies have focused on the CEFR, even though the word 'impact' sometimes appears in titles (e.g., Figueras, 2013; Jones & Saville, 2009; North, 2010). To date, no studies have used a Quality Assurance (QA) approach to identify which CEFR practices and which aspects of its implementation yield significant improvement in proficiency results. The lack of rigorous QA, feasibility and impact studies has led to very limited knowledge and expertise transfer between and within countries. This has resulted in inconsistencies in the implementation of what is, after all, a complex and comprehensive document, which in turn has reduced the benefits of the CEFR in terms of innovation in language policies and pedagogy.

QualiCEFR is the first international comparative study about CEFR implementation practices and the first to employ a QA process as its methodological approach. The project aims to: 1) identify and build on successful CEFR implementation strategies in different contexts; 2) facilitate the transfer of this knowledge and know-how in CEFR-related matters; 3) provide a template of principled guidelines that can be used by different stakeholders in implementing the

CEFR at the levels of policy, curriculum development, and teaching; 4) promote and facilitate a culture of evidence-based QA in CEFR implementation; 5) encourage reflection about key challenges and gains for language education that the CEFR offers.

## 2 Context of the research

QualiCEFR compares Switzerland and Canada, two multilingual countries with decentralized educational systems and high immigration rates. Switzerland and Canada are homes to linguistically diverse populations. In Switzerland, 59.8% of the population of just over 8 million speak a Swiss-German dialect as their mother tongue, and learn Standard German at school, 10.4% speak Standard German at home; 23.4% speak French, 8.4% Italian, and 0.6% Rätoromansch, the fourth national language. More than 23% of the Swiss population speak another language at home (Swiss Federal Statistics, 2017). Switzerland is very decentralized: each of the 26 cantons has sovereignty over its educational system. Coordination is assured by a Standing Conference of Education Directors (EDK), in Berne, the capital city. With all this diversity, it is not surprising that the initiative to develop the CEFR came from Switzerland, with the goal of creating a transparent and coherent system (Council of Europe, 1992).

Canada has two official languages: English and French. English is spoken by 57.8% of the population and French by 22.1%. Additionally, 20.1% of the Canadian population speak a language other than English or French at home (Statistics Canada, 2012). These non-official languages belong to many linguistic families and this diversity is increasing rapidly, given the immigration rate of more than 230,000 arriving annually since the early 1990s (Citizenship and Immigration Canada, 2016). Canada, like Switzerland, has no federal department of education nor an integrated national system of education. In the 10 provinces and three territories, departments or ministries of education are responsible for the organization, delivery, and assessment of education at all levels. The Council of Ministers of Education, Canada (CMEC) serves primarily as a means through which ministries can consult, but its role is not to steer educational policies at the pan-Canadian level. Nevertheless, in 2010, the CMEC officially embraced the recommendations for a common framework (Vandergrift, 2006), and published a document encouraging the use of the CEFR in Canada (CMEC, 2010). The CEFR has already informed K-12 language curricula revision in some provinces and knowledge mobilization around the CEFR is a core area of activity of the Canadian Association of Second Language Teachers (CASLT). Small-scale studies have suggested that Canadian language educators support the potential of the CEFR and that many of their priorities could be addressed by exploiting it (Faez, Majhanovich, Taylor, Smith, & Crowley, 2011; Mison & Jang, 2011; Piccardo, 2013; Arnott et al., 2017).

## 3 Theoretical framework

QualiCEFR builds upon QA expertise and draws on methods used in QA, program evaluation, and impact studies (e.g., Baird, 2007, Brown & Heyworth, 1999; Kiely & Rea-Dickins, 2005; Lasnier, Morfeld, North, Serra Borneto & Spaeth 2003; Martyniuk & Noyons, 2006; Matheidesz, 2010; Sheldon, 1987; Wall & Horák, 2011) to investigate and inform the implementation process of the CEFR. By doing this it aims to provide a unique and innovative methodology that will be available for potential replication on a wider scale.

Quality Assurance (QA) was first developed in industry and involves a systematic study of the design and production processes with each step in each process being defined with appropriate standards. The extent to which a product/service meets these standards is then inspected and evaluated. QA evolved further and incorporated two aspects: 1) human factors such as the encouragement of individual responsibility, team-work, job rotation, and 'quality circles' to encourage constant improvement; and 2) the extent to which the product/service fulfills the needs of different clients (Feigenbaum, 1951; 2015). QA has been applied in language education for analyses of curricula, resources, and processes, thus ensuring that humanistic, interpersonal, cognitive, and affective factors are at the centre of an inspection of planning, teaching, and assessment practices (Muresan, Heyworth, Mateva, & Rose, 2007; Heyworth, 2013). A typical QA scheme in language education may include analysis of documents and resources, interviews with key staff, classroom observations, focus groups with teachers and students, self-assessment questionnaires, and systematic evaluation applying the relevant quality standards (Matheidesz, 2010).

Another approach to the investigation of quality in education is program evaluation and the related concept of impact studies. The program evaluation approach remains relevant today (Kiely & Rea-Dickins, 2005), but impact studies have become more common, particularly in relation to desirable and undesirable effects on teaching caused by tests (washback). A typical impact study in language education may include analyzing curriculum documents/textbooks/tests, interviewing staff/teachers/students, and conducting surveys. One innovative project in the language context combined program evaluation and impact study methodology with a QA approach (Lasnier, Morfeld, North, Serra Borneto, & Spaeth, 2003), and produced a set of indicators for evaluating a book, a piece of software, or a course.

The choice to draw upon a theoretical framework that aligns with QA-informed approaches is linked to the lack of this type of study and to the appropriateness of this framework for the implementation of the CEFR, which is usually high stakes in terms of curriculum design, pedagogical and assessment reform.

## 4 Research methods used

The project adopts a multiphase mixed methods research design (Creswell & Plano Clark, 2011), with a sequential collection of qualitative data followed by quantitative data. Three types of data collection are being used: document search and semi-structured interviews (qualitative data), and surveys (quantitative data). This choice aligns with a design-based research paradigm as it combines empirical research with theory-driven design through the development of tools and the collaboration between researchers and practitioners (Design-Based Research Collective, 2003; Van den Akker, Gravemeijer, McKenney, & Nieveen, 2006, Anderson, & Shattuck, 2012). Rather than following a top-down approach, QualiCEFR is dependent on stakeholders' input, and the output (template) of this research will be shaped by their articulated needs.

The research includes two phases: 1) a comparative phase for mapping out CEFR-related initiatives to identify which ones have been effective and could be replicated and upscaled; 2) a QA phase for developing a CEFR-implementation template which will include QA procedures and indicators that can be used by education stakeholders. The template development will be informed by the results of Phase 1, and will highlight aspects of CEFR implementation that suggest high leverage in relation to the improvement of language proficiency.

The comparative phase started in September 2015 with a systematic search of published documents (online/hardcopy) from both countries to gather qualitative data about the relevance of the CEFR in second/foreign language education in Canada and Switzerland. Canadian documents dating from 2006, when the CEFR was first proposed in Canada (Vandergrift, 2006), and Swiss documents dating from 1991, the year when the CEFR project was initiated, have been considered. These documents included curricula, development projects, classroom projects, textbooks, and assessment procedures written in one of the official languages of the two countries. The focus has been on existing and/or absent CEFR-implementation initiatives. During the document analysis, emerging quality indicators of CEFR-related documents were identified. This analysis informed potential QA indicators to be further investigated through the second set of data collection: the semi-structured interviews.

The second set of data was collected through one-hour, semi-structured interviews with stakeholders in Canada (28 interviews) and in Switzerland (16 interviews). Stakeholders included representatives from educational authorities, language associations, CEFR-related project developers, and public-sector language education providers. The interviews provided answers regarding procedures (if any) that were used to both introduce and verify CEFR implementation, in order to identify successful strategies and procedures.

The interviews were then transcribed and member-checked. Subsequently, they were analyzed qualitatively using NVivo software to identify relevant, recurrent themes and related practices and procedures. From the thematic analysis, a granular taxonomy has emerged which helped the identification of effective practices and any QA procedures.

The first phase is being completed by collecting data quantitatively through an online survey that will be made available on SurveyMonkey.com for a two-month period. This survey, informed by the interviews previously conducted, includes questions to further investigate the QA procedures and techniques that emerged from the interviews in order to produce indicators for the QA template.

Once all the data has been analyzed, the team will move to Phase 2: conceptualization and development of the QA template. The template will be organized in sections for Program Design, Implementation and Evaluating Outcomes, with indicators that are cross-referenced to underlying quality principles (e.g., Reliability, Transparency, etc.). The content of the indicators will be guided principally by the analyses of the interviews and the survey from the previous phase. The template will be designed with a dual function in mind: (a) self-assessment as a tool for awareness-raising before or during an implementation, and (b) self- or external evaluation after an implementation. Stakeholders interviewed in Phase 1 will be invited to review drafts of the template during its development and to conduct a self-evaluation to pilot the final version.

## 5 Provisional results

Although the analysis is not yet completed, there are already some clear trends that can be identified. In both countries teacher education and the establishment of collaborative groups to foster innovation are major vectors of change. As one Canadian participant put it: “It is important to engage in ongoing learning … not just one workshop but regular reconnection with support on how to take the implementation further”. Interestingly, a Swiss participant reported that “a taskforce was launched to develop both the capacity among teaching staff AND a bank of resources to help them teach and evaluate according to this new approach”. In Switzerland, 63% mention teacher education as a successful strategy, with 44% stating positive impact, while the figures for Canada are 61% and 54%. In Switzerland 69% mention collaborative groups as a successful strategy, though only 19% cite positive impact; in Canada 43% cite such groups as a successful strategy with the same proportion reporting positive impact. Respondents in both countries cite language portfolios as a vector, with 44% in Switzerland citing positive impact, despite admitted signs of ‘portfolio fatigue.’

The biggest difference between the two countries is the strategy of introducing an external CEFR-based examination to engender change. In Switzerland experience with this strategy appears mixed, whereas in Canada 46% report the implementation of the DELF as a positive strategy, with 39% citing positive impact. In fact, in Canada participants reported that “the DELF is the catalyst, the vehicle of implementation of the CEFR, because once you get that up and happening, it trickles down to levels, instructional practices, etc.”

In terms of overall approach, a number of respondents in both countries stressed the importance of involving all stakeholders (especially the political administration) and adopting multiple strategies rather than relying on a single approach. As participants eloquently stressed “Every level of the organization has to be playing their role in it … It has to be a full-on” and “Everyone needs to see themselves in this picture”. A combination of top-down and bottom-up initiatives was mentioned as the ideal scenario by respondents in both countries, with those in Canada lamenting that nearly all initiatives were just bottom-up with little support from the authorities, and those in Switzerland lamenting that most CEFR-related initiatives had been solely top-down. Those initiatives that had been successful tended to be those that had planned out a project in detail with defined stages, organized scientific accompaniment, recruited a large number of teachers early in the development process, and delivered a concrete product that addressed a felt need.

## 6 Conclusion

Curriculum documents do not achieve change per se; change is achieved through planning an appropriate implementation and by providing a means to ensure that curriculum principles are followed and developed in practice. By introducing a QA approach to CEFR implementation, QualiCEFR aims to systematically enhance coherence between objectives, planning, and practice in the long term. QualiCEFR is also encouraging cross-fertilization of ideas among Canadian and European researchers, by providing a model that can help bridge the knowledge transfer gap between countries.

Through its QA-informed template, QualiCEFR aims to provide policy makers and stakeholders with critical knowledge and tools necessary to successfully implement the CEFR in their own contexts. Such a QA approach should benefit language policies in Canadian contexts that are considering adopting the CEFR and/or are in the initial stages of implementation, and it will potentially enhance professional practices among teachers to initiate a radical shift in pedagogy.

QualiCEFR will help raise awareness about the need to exchange knowledge and know-how between contexts with a different level of experience in the CEFR, and to follow QA approaches in the implementation of the CEFR. To date, no study of CEFR implementation has followed such a QA approach: QualiCEFR will hopefully act as a catalyst for research in the academic sector and in educational institutions, by encouraging studies about CEFR implementation and its impact.

QualiCEFR should ultimately benefit the general public in our increasingly multilingual societies by providing a teacher-friendly template to help practitioners understand how to innovate in their contexts, improve learner motivation and proficiency, and achieve that real-world second language ability originally envisioned by the CEFR.

## References

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher, 41*(1), 16–25.

Arnott, S., Brogden, L., Faez, F., Péguret, M., Piccardo, E., Rehner, K., Taylor, S., & Wernicke, M. (2017). The Common European Framework of Reference (CEFR) in Canada: A Research Agenda. *The Canadian Journal of Applied Linguistics (CJAL) 20*(1), 31–54.

Baird, J.-A. (2007). Investigating Washback in Language Testing and Assessment, Special Issue: *Assessment in Education: Principles, Policy & Practice, 14*(1), 1–137.

Brown, P. & Heyworth, F. (1999). *Modern languages: Learning, teaching, assessment. A Common European Framework of Reference: A user guide for quality assurance and quality control*. Strasbourg, France: Council of Europe, DECS/EDU/LANG (99) 17.

Byram, M. & Parmenter, L. (Eds.). (2012). *The Common European Framework of Reference: The globalisation of language policy*. Bristol, UK: Multilingual Matters.

Citizenship and Immigration Canada. (2016). *150 years of immigration in Canada*. Retrieved from http://www.statcan.gc.ca/pub/11-630-x/11-630-x2016006-eng.htm

Council of Europe. (1992). *Transparency and Coherence in Language Learning in Europe: Objectives, assessment and certification: the proceedings of the Intergovernmental Symposium held at Rüschlikon November 1991*. Report by B. North, Strasbourg: Council of Europe.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Ministers of Education, Canada. (2010). *Working with the Common European Framework of Reference for Languages (CEFR) in the Canadian context: Guide for policy-makers and curriculum designers*. Retrieved from https://www.cmec.ca/docs/assessment/CEFR-canadian-context.pdf

Creswell, J. W. & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Design-Based Research Collective. (2003). *Design-based research: An emerging paradigm for educational inquiry. Educational Researcher, 32*(1), 5–8.

Faez, F., Majhanovich, S., Taylor, S., Smith, M., & Crowley, K. (2011). The power of "can do" statements: Teachers' perceptions of CEFR-informed instruction in French as a second language classrooms in Ontario. *The Canadian Journal of Applied Linguistics, 14*(2), 1–19.

Feigenbaum, A. (1951/2015). *Total Quality Control*. New York: McGraw Hill.

Figueras, N. (2013). The impact of the CEFR. *English Language Teaching Journal, 66*(4), 77–485.

Heyworth, F. (2013). Applications of quality management in language education. *Language Teaching, 48*(3), 281–315.

Jones, N. & Saville, N. (2009). European language policy: assessment, learning and the CEFR, *Annual Review of Applied Linguistics, 29*, 51–63.

Kiely, R. & Rea-Dickins, P. (2005). *Program Evaluation in Language Education*. London: Palgrave Macmillan.

Lasnier, J-C., Morfeld, P., North, B., Serra Borneto, C., & Spaeth, P., (2003). *A Guide for the Evaluation and Design of Quality Language Learning and Teaching Programmes and Materials* [CD-ROM]. A project related to the WHITE PAPER "Teaching and learning. Towards the learning society" Objective 4, 1st support measure. Co-funded by DG XXII, EUROPEAN COMMISSION, Brussels.

Martyniuk, W. & Noyons, J. (2006). *Executive summary of results of a survey on the use of the CEFR at national level in the Council of Europe Member States, 2006*, Strasbourg, France: Council of Europe.

Matheidesz, M. (2010). International accreditation of quality in language learning and teaching. *Research Notes, 39*, 33–39.

Mison, S., & Jang, I. C. (2011). Canadian FSL teachers' assessment practices and needs: implications for the adoption of the CEFR in a Canadian context. *Synergies Europe, 6*, 99–108. Retrieved from https://gerflint.fr/Base/Europe6/Europe6.html

Muresan, L., Heyworth, F., Mateva, G., & Rose, M. (2007). *Qualitraining: A training guide for quality assurance in language education*. Strasbourg: Council of Europe.

North, B. (2010). The Educational and Social Impact of the CEFR. In L. Taylor & C. Weir (Eds.), *Language Testing Matters: Investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge Conference, April 2008, Cambridge*. Studies in Language Testing series 31, Cambridge, UK: Cambridge University Press, 357–377.

Piccardo, E. (2013). (Re)conceptualiser l'enseignement d'une langue seconde à l'aide d'outils d'évaluations: comment les enseignants canadiens perçoivent le CECR. *The Canadian Modern Language Review/La Revue Canadienne des langues vivantes (CMLR/RCLV) 69*(4), 386–414.

Sheldon, L. E. (Ed.). (1987). *ELT textbooks and materials: Problems in evaluation and development* (ELT Document 126; pp. 37–44). London, UK: Modern English Publications in association with the British Council.

Statistics Canada. (2012). *Linguistic Characteristics of Canadians*. Retrieved from http://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011001-eng.cfm

Swiss Federal Statistics. (2017). *Langues les plus fréquentes parlées habituellement à la maison par la population résidente permanente âgée de 15 ans ou plus*. Retrieved from https://www.bfs.admin.ch/bfs/fr/home/statistiques/population/langues-religions.assetdetail.2263148.html

Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (2006). *Educational Design Research*. London: Routledge.

Vandergrift, L. (2006). *New Canadian perspectives: Proposal for a common framework of reference for languages in Canada*. Department of Canadian Heritage. Retrieved from https://www.caslt.org/pdf/Proposal_Common%20Framework_Reference_languages%20for%20Canada_PDF_Internet_e.pdf

Wall, D. & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2: Coping with change*. TOEFL iBT Research Report, TOEFL iBT–5, July 2008, Princeton: Educational Testing Service.

# Tatar Language Tests for Maintaining a Language of a Minority Ethnic Group in Russia: The Case of Kazan

**Marina Ivanovna Solnyshkina**[6], Leo Tolstoy Institute of Philology and Intercultural Communication, Kazan (Volga Region) Federal University, Russia
**Gulnara Vasilevna Sadykova**, Leo Tolstoy Institute of Philology and Intercultural Communication, Kazan (Volga Region) Federal University, Russia
**Alsu Khaliloavna Ashrapova**, Leo Tolstoy Institute of Philology and Intercultural Communication, Kazan (Volga Region) Federal University, Russia
**Elena Vladimirovna Harkova**, Leo Tolstoy Institute of Philology and Intercultural Communication, Kazan (Volga Region) Federal University, Russia

**Abstract:** The Tatar language tests developed on CEFR ideas is a relatively new phenomenon. The impact of these tests on language education in the Republic of Tatarstan and their function to serve as a tool to maintain a language and transfer cultural values have not yet been studied. Addressing this gap, the paper reports on the survey held to determine whether Tatar tests' structure and content meet the requirements of the target audience, i.e. graduates of the course Tatar as a Second/Foreign Language offered by Kazan Federal University (KFU) for the general public. To assess the graduates' needs, two questionnaires were developed and administrated in Leo Tolstoy Institute of Philology and Intercultural Communication at KFU in the autumns of 2015 and 2016 (N=457). The research addressed the graduates' views on the most required topics and domains where they plan to use the Tatar language to succeed in their life and studies. The content analysis of the responses showed a broad variety of needs (from reading street signs to taking a university course), although the most frequent demands mentioned in over 74% of the questionnaires proved to be the acquisition of Tatar culture and traditions. The findings also show that the 2015 group of Tatar course students found that the reading and listening tests were unchallenging and lacked a distinctive ethnic identity. Based on the questionnaires' analysis the authors of the article advocate reconceptualization of the Tatar language course curriculum as a response to the graduates' expectations and the modern Tatar language contexts. The latter must lead to changes in the language material selection both for teaching and testing so that the course participants' needs and expectations would be adequately met and the domains where Tatar is used extended.

## 1 Introduction

In the professional community of teachers of Tatar, communicative purposes of language learning were declared long ago but began being recognized and implemented by in-service teachers as late as in the 1990s (Guzelbaeva & Fatkhullova, 2012). Since then, the Republic of Tatarstan (RT) has been witnessing a Tatar language learning boom accompanied by educators developing new courses and programmes, publishers lining up for Tatar language textbooks and dictionary copyrights (Gimaletdinova & Khalitova, 2016). Since then there have been some important developments in the Tatar language curriculum in Tatarstan including the introduction of state exams for secondary and high school graduates, the Republic Main Exam (RME) and the Republic Unified Exam (RUE), correspondingly. The common efforts of the government, media, the system of education and cultural activists led to significant progress in extending the Tatar language domains, but the quality of Tatar language instruction still leaves much to be desired (Yusupov, Aydarova, Sagdieva & Harisova, 2015). Addressing the problem, in 2013, the Ministry of Education and Science of RT teamed up with Kazan Federal University and

[6] This research was supported by a subsidy of the Russian Government to support the Program of Competitive Growth of Kazan Federal University.

Correspondence concerning this article should be addressed to Marina I. Solnyshkina, Department of Germanic Philology, Leo Tolstoy Institute of Philology and Intercultural Communication, Kazan (Volga Region) Federal University, Kazan, ul.Kremlevskaya, 18, Kazan, 420008, Russian Federation. Email: mesoln@yandex.ru

introduced a free Tatar language course for the general public. The complex aim of the course was to disseminate Tatar traditions and improve learners' understanding of the Tatar culture. As the course gained popularity, it was decided to design, develop and conduct Tatar language tests based on the principles of the Common European Framework of Reference for Languages (CEFR). The pilot tests held in 2015 in Leo Tolstoy Institute of Philology and Intercultural Communication proved the tasks to be too simple as over 80% of the test-takers received maximum scores. Afterwards the authors of the article developed two Tatar language test questionnaires and administered them in 2015 and 2016 to all the students of the course (N=457).

This paper presents preliminary results of the Tatar language needs analysis aimed at identifying how the information collected can be used to design the Tatar language courses to better serve the learners' needs and thus extend the Tatar language domains in RT. The research focuses on the following questions: What is the social context in which the Tatar tests have been developed? What Tatar language testing is needed? What must be the content of Tatar language tests?

## 2 Literature review

In large multicultural nations, the maintenance of a minority language requires significant efforts (Nettle & Romaine, 2000). These efforts should be doubled or tripled in nations where one dominant language is established as a single state language and as a language of formal education (Dorian, 1982). Such is the case of the Tatar language that exists within the Russian language environment in the Russian Federation as a titular language on the one hand and a language of a minority ethnic group on the other.

Tatars are traditionally labelled as a one of the 'titular' nations in the Russian Federation, meaning that the name of this ethnic group is given to the corresponding subject of the federation – the Republic of Tatarstan (Malakhov & Osipov, 2006, p. 504). But Tatars also make a minority ethnic group in the Russian Federation since Russian as a dominant language is used "in most official domains: government, public offices, and educational institutions" (p. 9) and the minority language (i.e. Tatar) continues "to be integral to a number of public domains, especially in traditional religious institutions, local stores, and those places where members of the community socialize" (p. 9). The authors of *Language Vitality and Endangerment* (UNESCO Ad Hoc Expert Group on Endangered Languages, 2003) specify the situation as follows:

> The described coexistence of languages results in speakers' using each language for a different function (diglossia), whereby the non-dominant language is used in informal and home contexts and the dominant language is used in official and public contexts. Speakers may consider the dominant language to be the language of social and economic opportunity. However, older members of the community may continue to use only their own minority language. (p. 9)

Though Tatars remain the second largest ethnic group in the Russian Federation (after Russians) and despite the fact that over 53% of the population in the Republic of Tatarstan are ethnic Tatars (Federalnaja Sluzhba Gosudarstvennoj Statistiki, 2010), the issue of preserving and maintaining the Tatar language is one of the most urgent in the agenda of Tatarstan policy makers (Garipov & Solnyshkina, 2004). One of the reasons for this is the ongoing downturn of the Tatar-speaking population: the 2010 all-Russia census registered a 19.95% decline of people who claimed to speak Tatar in comparison with the number in 2002 (Federalnaja Sluzhba Gosudarstvennoj Statistiki, 2002; 2010).

## 3 Social context of the test development

Historically Tatars lived side-by-side with Russians, but due to the language policies in the Soviet Union and global changes, the scope of the Tatar language usage became limited and it gradually lost its prestige, especially among young Tatar citizens. By the late 1980s linguists recorded the situation of diglossia with Russian holding a high social status and Tatar being a means of everyday communication in informal contexts (Guzelbaeva & Fatkhullova, 2012).

The rise of national self-identity in the early 1990s led to the gradual revival of the Tatar language. While to date, the asymmetry of Russian–Tatar usage has not been eliminated (ibid.), the Decree of 1992 (Natcionalnaja Biblioteka Respubliki Tatarstan, 2014) established the equality of Russian and Tatar as two state languages in the Republic of Tatarstan, which enabled the local government to expand the contexts and increase the amount of the Tatar language usage in the society. At present, it is obligatory for all children in RT to attend Tatar language classes and a Tatar language exam is a requirement for all secondary and high school graduates. In RT, the Tatar language and literature classes in grades 1–9 have the same proportion of study hours as the Russian language and literature classes. Tatar language command is also a requirement for a number of jobs in the public sector. The Tatarstan government invests in the development of Tatar TV and radio channels, journals and magazines (Garipov & Solnyshkina, 2006).

Even though the Law of 1992 has partly mitigated the diglossia effect in Tatarstan (Law on Languages, 1992), research, censuses and mass media indicate that Tatar still remains the language of informal communication of mostly Tatars. The majority of Russians in Tatarstan are not able to communicate in Tatar. According to Akhmetova, Guzelbaeva, Eflova, Nizamova & Nurutdinova (2012), 63% of the population in Tatarstan understand Tatar and 53% speak colloquial Tatar, 86% of Tatars “speak it well” and only 4% of Tatars do not speak it at all; only 7.5% of Russians in Tatarstan speak Tatar and 18% of Russians in RT understand Tatar. Young people in RT are much more motivated to study English than Tatar (Pravda, n.d.).

The mandatory classes of Tatar language and literature in all RT public schools have not yet changed the situation radically. Guzelbaeva & Fatkhullova (2012) explain it by the gaps in Tatar education. They claim that schools need to (1) focus on communicative real-world skills; (2) develop teaching skills of Tatar teachers, involve ICT and new methods of teaching; (3) improve existing coursebooks; (4) develop a series of coursebooks that will target learners from preschools to graduate schools.

The situation is complicated by the negative attitude of some parents who do not see the need for their children to study Tatar on a par with Russian (Sulejmanov, 2014). In 2011–2012 a group of parents wrote an open letter to the Minister of Education and Science of the Russian Federation claiming that Tatar is taught at the cost of Russian classes and, thus, limits students' time spent on studying Russian and preparing for high-stakes exams. The latter, however, demonstrated that there was no significant difference in the exam results of those who had taken Tatar classes and those who had not (Guzelbaeva & Fatkhullova, 2012).

When in 2012 the RT Ministry of Education and Science funded the first Tatar language course for the general public, over 1200 people from Kazan, the capital of Tatarstan, signed up for the evening classes. At that point there were no tests that would enable the Ministry (and other RT stakeholders at large) to obtain an objective evaluation of Tatar language proficiency. In 2014 the RT Ministry of Education and Science launched the State Programme for Maintaining, Studying and Developing the Languages of the Republic of Tatarstan in 2014–2020 (Elektonnyj, n.d.), which directly affected the usage of Tatar in RT. As part of implementing the State Programme, Kazan Federal University, the only institution of higher education in the Russian Federation that offers bachelor, master and doctoral programmes in the Tatar language, literature and culture, established the Tatar Language Assessment Centre. The Centre united over 30 linguists and testologists of Kazan Federal University with the aim to develop a robust system of assessment of language proficiency. By now the Tatar Language Assessment Centre had designed, developed and piloted two batteries of A1–B2 Tatar tests.

At the beginning of the process in 2012, KFU test developers had to address at least two major concerns. Firstly, they had to target a heterogeneous group of learners, some of whom possessed receptive skills only (learning and reading). Secondly, there was a lack of understanding among test developers on the content to be assessed in Tatar tests.

KFU educators designed the curriculum for Tatar as a Second/Foreign Language course as part of implementing the State Programme. As mentioned above, the course serves the two-fold social purpose: to disseminate Tatar traditions and improve learners' understanding of Tatar culture. The developers ambitiously intended to create a tailor-made course designed on learners' "necessities, lacks and wants" (Nation & Macalister, 2010).

## 4 Methods and instruments

The main method employed in the research presented is needs analysis viewed by the authors as "an array of procedures that can identify, validate, and prioritize needs" (Pratt, 1980). Needs analysis was used to assess whether the previously specified and implemented in KFU educational goals correspond to the Tatar language course participants' "wants" or "desires". Following and developing Widdowson's ideas (1981) we define needs as the requirements/felt needs/objectives of learners on the content and goals of the language course taken, thus sharing the goal-oriented approach to the notion. We also assume the necessity in a number of situations to study "the gap between the current situation and the anticipated future state" (Berwick, 1989, p. 52) or linguistic "inadequacies to be filled" (Robinson, 1991).

Thus, the research is mostly focused on the students' "target needs" as "necessities", "lacks" and "wants" for being successful "in the target situation" (Hutchinson & Waters, 1987). The latter is the situation in which a student plans to function effectively. For example, for Tatar language course students it is important to master their listening skills to be able to watch a Tatar theatre performance without headphones.

The instruments selected for the research were questionnaires which consisted of a series of questions aimed at identifying the course participants' "subjective needs" (as defined by Graves, 2000), i.e. "underlying purposes".

The questionnaires were distributed by the method of direct contact after the completion of the course when the participants received their scores. The survey was anonymous; the time limited to answer the questions was 15 minutes.

The questions used by the authors were classified in three groups: (1) closed, (2) open response-option questions and (3) open-ended.

The closed type of questions referred mostly to the participants' demographic information: age, gender, occupation, language proficiency and test structure.

The demographic dimensions of the participants are as follows:

**Age groups**

| Age | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 | 65+ |
|---|---|---|---|---|---|---|
| % | 42 | 25 | 14 | 8 | 6 | 5 |

**Gender**

| Gender | Male | Female |
|---|---|---|
| % | 27 | 73 |

**Occupation**

| Occupation | Students | Teachers | Business | Retired | Musicians | Health services | Designers | Accountants | Workers |
|---|---|---|---|---|---|---|---|---|---|
| % | 37 | 20 | 19 | 12 | 1 | 1 | 1 | 1 | 1 |

**Language proficiency**

| Level | A1 | A2 | B1 | B2 |
|---|---|---|---|---|
| % | 67 | 15 | 11 | 7 |

**Table 1.** Demographic information about the participants

The course participants unanimously supported the 5-part structure of Tatar tests containing Reading, Listening, Writing, Grammar and Vocabulary and Speaking parts.

## 5 What Tatar language testing is needed?

The closed type questions referred to the communicative needs of the learners in the target situation. The communicative needs are viewed by the authors broadly as the setting in which the learners plan to use the target language, the learners' role in the target situation, necessary language skills, language proficiency required by the target situation (Hutchinson & Waters,1987; Richards, 1990).

The first question on communicative needs was presented to the informants as follows: "1. Please, select and tick one or more language skills you consider as the most significant." The options included Speaking, Reading, Writing, Listening, Grammar, Vocabulary. As it was anticipated, Speaking was marked by 100% of participants, thus proving to be perceived as the most important skill. Vocabulary was selected by 50% of the participants, revealing the shared view on its importance in target communicative situations. The figures for Listening (47%), Vocabulary (50%) and Grammar (40%) were consistently lower than those for Writing (22%) and Reading (17%).

Open response-option questions were used to identify the language domains, i.e. the situations where the learners intend to use Tatar. We provided them with the following options: "with your family members, neighbours, shop assistants, friends, elderly people, schoolmates/colleagues/fellow students, in hospitals, at conferences, over the phone, at work, reading for entertainment, expressing emotions, in classes of Tatar, other." We intentionally presented a number of overlapping options, thus eliciting the information which otherwise could have been omitted. The answers on the language skills were later on correlated with the domains selected by the graduates, thus providing us with the foundation to verify the responses collected.

As it might be expected, family life was the domain marked by the majority of the participants: 87% of the 457 stated that they study Tatar to speak at home (1st position). Surprisingly, the "work" domain gained over 35% of choices, thus certifying the popularity of Tatar in a business environment (2nd position). Theatre (10%), concerts (7%) and news (4%) make together over 20% (3rd position) and are higher than that for educational purposes (12%). 100% of hospitality business employees as well as doctors and nurses selected Tatar as the language to be spoken with customers/patients and thus brought this domain to fourth position with 14%. It can also serve as a sign of the social climate in RT to be more conducive to the use of Tatar than it was a decade ago (Wertheim, 2002).

## 6 What must be the content of Tatar language tests?

The questions of an open type were asked to seek the information on participants' content needs which include the selection and arrangement of topics, grammar, vocabulary, language functions and notions (Nunan, 1999).

The findings show that students are inclined to view the Reading and Listening tasks as lacking ethnic markedness, boring and deprived of cognitive challenges. The course graduates also generated a list of mandatory topics which includes (in decreasing order): everyday life in Tatarstan (99%), Tatar artists and Arts (92%), Tatar traditions (86%), Tatar holidays (83%), Tatar literature (78%), Tatar cuisine (40%), Tatar music (34%), Tatar ethnic outfit (22%). The figures prove that the participants regard Tatar as a tool for transmission of culture and values.

## 7 Conclusion

The needs analysis conducted revealed a range of Tatar course graduates' learning needs, which were based on their aspirations, career plans and motivation. It proved that though Tatar is still predominantly required for family communication, cultural transmission and educational purposes, there is a well observed demand for Tatar for specific purposes in business and health services. The ongoing language shift in the domains where Tatar functions and the gradual extent of the domains reveal a favourable attitude of the RT citizens towards Tatar.

The research also exposed the conflict of the language needs of the graduates with the existing course content which was characterized as demotivating and boring.

Based on the results of the needs analysis pursued, KFU educators are designing the Tatar language test specifications (CEFR A1–B2) that address cognitive, linguistic and affective needs of the target audience. The new curriculum is planned to be recommended for practice in KFU in the autumn of 2017.

## References

Akhmetova, S. A., Guzelbaeva, G. Y., Eflova, M. Y., Nizamova, L. R., & Nurutdinova, A. N. (2012). *Sostojanie i dinamika mezhetnicheskih I mezhkonfessionalnyh otnoshenij v Respublike Tatarstan*. Kazan, Russia: Kazan Federal University.

Berwick, R. (1989). Needs assessment in language programming: from theory to practice. In R. Johnson (Ed.), *The Second Language Curriculum* (pp. 48–62). Cambridge: Cambridge University Press.

Dorian, N. C. (1982). Language loss and maintenance in language contact situations. In R. D. Lambert & B. F. Freed (Eds.), *The Loss of Language Skills* (pp. 44–59). Rowley, MA: Newbury House.

Elektonnyj Fond. (n.d.). *Ob utverzhdenii gosudarstvennoj programmy Sokhranenie, isychenie n razvitie gosudarstvennykh jazykov Respubliki Tatarstan I drugikh jazykov v Respublike Tatarstan na 2014-2020 gody* [Official Document]. Retrieved from http://docs.cntd.ru/document/463305579

Federalnaja Sluzhba Gosudarstvennoj Statistiki. (2002). *Vserosijskaja Perepis Naselenia 2002.* [Data File]. Retrieved from http://www.perepis2002.ru/index.html?id=11

Federalnaja Sluzhba Gosudarstvennoj Statistiki. (2010). *Vserosijskaja Perepis Naselenia 2010*. [Data File]. Retrieved from http://www.gks.ru/free_doc/new_site/perepis2010/croc/perepis_ itogi1612.htm

Garipov, Y., & Solnyshkina, M. (2004). *Language Policy in National Republics of Russia: Ideology and Practice of Tatarstan.* Language and the Future of Europe: Ideologies, Policies and Practices, University of Southampton, Southampton, UK. Retrieved from http://www.lang.soton.ac.uk /lipp/abstracts.html

Garipov, Y. & Solnyshkina, M. (2006). Language Reforms in Tatarstan Education System and the Ethnolingustic Orientation of Young People. *Voices Diversae: Lesser Used Language Education in Europe*. Belfast, Northern Ireland: Clo Ollscoil na Banriona, 131–138.

Gimaletdinova, G., & Khalitova, L. (2016). Self-paced learning: Investigating an online Tatar language course. *XLinguae Journal, 9*(3), 81–92.

Graves, K. (2000). A framework of course development processes. In D.R. Hall & A. Hewings (Eds.), *Innovation in English language teaching* (pp. 179-196). London: Routledge.

Guzelbaeva, G. Y., & Fatkhullova, K. S. (2012). Relizatciya jazykovoj politiki i puti vyravnivanija jazykovoj assimetrii v Respubliki Tatarstan. *Phylology & Culture, 3*(29), 35–41.

Hutchinson, T., & Waters, A. (1987). *English for Specific Purposes: A learner-centered approach*. Cambridge University Press.

Law on Languages. (1992). *Vedomosti Verkhovnogo Soveta Tatarstana, 6*, 3–10.

Malakhov, V., & Osipov, A. (2006). The Category of “Minorities” in the Russian Federation: Reflection on Uses and Misuses. In Sia Spiliopoulou Akermark (Ed.), *International Obligations and National Debates: Minorities around the Baltic Sea*. Mariehamn, Aland, Finland: Alands Islands Peace Research Institute, pp. 497–544.

Natcionalnaja Biblioteka Respubliki Tatarstan. (2014). *Zakon Respubliki Tatarstan ot 08.07.1992 N1560-XII O gosudarstvennykh jazykakh Respubliki Tatarstan i o drugikh jazykakh Respubliki Tatarstan* [Official Document]. Retrieved from http://kitaphane.tatarstan.ru/legal_info/zrt/lang.htm

Nation, I. S. P., & Macalister, J. (2010). *Language Curriculum Design*. New York, NY: Routledge.

Nettle, D., & Romaine, S. (2000). *Vanishing Voices: The Extinction of the World’s Languages*. Oxford: Oxford University Press.

Nunan, D. (1999). *Second Language Teaching and Learning*. Boston, MA: Heinle & Heinle Publishers.

Pratt, D. (1980). *Curriculum design and development*. New York, NY: Harcourt Brace.

Pravda li, chto tatarskij ischezaet? (n.d.). Kazan.ru. Retrieved from https://inkazan.ru/news/society/21-02-2017/pravda-li-chto-tatarskiy-yazyk-ischezaet

Richards, J. C. (1990). *The language teaching matrix*. Cambridge: Cambridge University Press.

Robinson, P. (1991). *ESP Today: A Practitioner’s Guide*. Hemel Hempstead: Prentice Hall.

Sulejmanov, R. R. (2014). *Etnolingvisticheskij konflikt v sovremennom Tatarstane: Borba za russkijjazyk v shkolakh natcionalnoj respubliki*. Proceedings from Jazakovaja Politika I Jazykovyje Konflikty v Sovremennom Mire (Moscow, September 16–19, 2014), 226–234.

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). *Language Vitality and Endangerment*. Retrieved from http://www.unesco.org/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/unesco_language_vitaly_and_endangerment_methodological_guideline.pdf.

Wertheim, S. (2002). *Language purity and the de-russification of Tatar*. Retrieved from http://iseees.berkeley.edu/sites/default/files/u4/bps_/publications_/2002_02-wert.pdf.

Widdowson, H. G. (1981). English for specific purposes: Criteria for course design. In Selinker et al. (Eds.), *English for Academic and Technical Purposes: studies in honor of Louis Trimble* (pp. 1–11). Rowley, MA: Newbury House.

Yusupov, R. A., Aydarova, S. H., Sagdieva, R. K., & Harisova, G. F. (2015). Improving efficiency of teaching the Tatar language to a foreign audience. *International Education Studies, 8*(5), 158–164.

# The Use of Test Taker Productions in Redesigning Writing Assessment Grids. A Corpus Based Study

**Dina Vîlcu**, Babeș-Bolyai University, Romania
**Antonela Arieșan**, Babeș-Bolyai University, Romania
**Lavinia Vasiu**, Babeș-Bolyai University, Romania

**Abstract:** Committed to permanently improving the quality of assessment at the Department of Romanian language, culture and civilisation (Babeș-Bolyai University, Cluj-Napoca), we are in a constant dialogue with assessors, experts in the domain and stakeholders, making their reactions and feedback part of the process. The present study reflects the process of transformation of our expert-designed writing assessment grid, focusing on the criterion of accuracy at level A1. Based on feedback from assessors and processed data from the corpus of our test takers' written productions, some of the resulting changes were: including new descriptors and abandoning the ones which proved irrelevant, introducing intermediate bands, eliminating imprecision of terms from descriptors, change of the order of criteria in the grid, and separate assessment of the tasks. The project will continue with reshaping of the other criteria in the grid, for levels A1–B2, and with the reconstruction of the grids for the spoken productions.

## 1 Introduction

As part of the reshaping of the assessment process at our department, the expert team developed grids for the evaluation of the written and spoken productions, for the levels A1–B2, in close connection with the CEFR. The whole testing process reflects the communicative competence as it is described in the CEFR, including not only linguistic knowledge, but also socio-linguistic, and pragmatic strategies, a perspective which is based on the model proposed by Canale and Swain (1980). Also the criteria we proposed for assessment follow the scales and the grids in the CEFR (Council of Europe, 2001).

The grids we have been using in the assessment process were developed about four years ago and in comparison with the previous rather unsystematic assessment, the use of the new grids has brought a tremendous improvement in terms of objectivity, consistency and fairness to the test takers. After a series of 15 successive evaluation sessions, feedback on the adequacy of the writing grids was collected. Thirteen assessors responded with feedback and they felt that the grids needed improvement in terms of: precision of terms used in descriptors (notions like syntactic structures, lexical means and sometimes, frequently or often in descriptors like "He/She sometimes makes orthography and punctuation mistakes" were pointed out as unclear and confusing); differentiation between the bands, with some descriptors identical from one band to the next; overlapping of criteria (simple sentences or short sentences are elements belonging to the criterion of complexity, not accuracy). There has also been positive feedback, the precision of some descriptors being appreciated by many of the assessors. Besides the direct feedback on the grids, the assessors also had suggestions related to the whole evaluation process. One of the most important suggestions was that of having at hand, while checking the productions, a minimal inventory of grammar forms and lexical means the candidates should be able to use at a certain level.This instrument was made available to the assessors, being, at the same time, constructed in relation with what is known as foreigner talk (Ferguson, 1971) and was described also as micro-language (Platon, 2016). In terms of actual content, it was related to the minimal description of Romanian language (Platon, Sonea, Vasiu, & Vîlcu, 2014) and the syllabus reflected in the textbook of Romanian as a foreign language produced at our department (Platon, Sonea, & Vîlcu, 2012). Other suggestions coming from the assessors were related to the possibility of using intermediate bands in the grids and separate assessment of the two tasks. This last suggestion was seen as an opportunity to solve certain problems which sometimes appear in relation to some of the productions, like the situations when the candidate only responds to one of the tasks. We considered that an informed reshaping of the grids and their criteria could only be made by use of input from test takers' productions. All these factors proved to correspond closely to the key questions Sara Cushing Weigle mentions we need to consider before designing assessment tasks and scoring procedures (Cushing Weigle, 2002). These questions are not only related to the content of the tasks, but, among others, also to who our test takers are, who will score the test and based on what criteria, and who will use the information that our test provides. We decided, thus, to reshape our grids, changing from a measurement-driven grid to a performance-driven one (Deygers, Van Gorp, Joos & Luyten, 2013) and taking into consideration the advantages of empirically developed rating scales in comparison with the intuitively developed ones (Knoch, 2002). The grids are, thus, based on the linguistic reality of those learning the language, on interlanguage at level A1 (Selinker, 1972). The presentation of our study will show the way in which we assembled and processed the corpus of student productions and the results we arrived at.

.

## 2 Assembling of corpus

The corpus includes, at present, 352 productions (16,190 words), coming from 126 test takers, who had previously agreed to the use of their productions for research. We intend to increase the number of productions in our corpus, making it and the research based on it more and more representative and relevant. The examinations were administered at our department between 2013 and 2016 and the tasks are of two types: writing an email (a presentation of a regular school day; a presentation of a free day) and description (description of a person, description of a room).

The criteria of selection applied for assembling the corpus were related to cut off and to representativeness in terms of mother tongue and gender. By use of these criteria we tried to meet McEnery and Hardie's recommendation: 'the corpus data we select to explore a research question must be well matched to that research question' (McEnery & Hardie, 2012). Thus, we selected productions which were over the cut off score (8 or 9 out of 20). The cut off score is calculated, using the method of contrasting groups, for each component of the examination and can vary between 8 and 12. We decided to exclude the productions which were below cut off score because the number of mistakes made large parts of those texts impossible to comprehend and irrelevant for characterising the test takers' ability to produce written text at level A1.

The consideration of the criterion of the mother tongue kept our corpus at a rather low number of productions. In order for the corpus to be balanced from the point of view of L1, we tried to select productions representing all the mother tongues of the candidates in a proportional way. However, every year most of the students participating in our courses and examination have Arabic as their mother tongue (in the academic year 2016–2017, the proportion of students with Arabic as a mother tongue was 75% – 88 out of a total of 115 students). In these conditions, our goal was to keep the number of productions coming from students with Arabic as a mother tongue at around one third of the total of productions. Thus, in terms of L1, our corpus of written productions for level A1 reflects, for the moment, the following distribution: 39% Arabic; 14% Albanian; 6% Italian; 6% French; 6% Spanish; 11% Asian languages (Chinese, Japanese, Thai); 5% African languages; and 13% other languages.

## 3 Corpus analysis

### 3.1 Elements considered

We started the corpus analysis from the elements which had been included in the old shape of the grid (grammar, vocabulary, orthography, and punctuation). However, the process of analysis took us to the conclusion that other elements also need to be considered (word order and language transfers – language transfers considered as strategies of communication) (Corder, 1983). The writing in capital letters also became part of a descriptor, together with the punctuation, and the notion of incoherent sequences (due to grammar or vocabulary mistakes) was introduced in the grid. The analysis was, thus, modelled in a very relevant way by the information we obtained from the corpus (introduction of new descriptors, combination of descriptors in one formula), which could have only been revealed by the actual productions of the candidates.

### 3.2 Method

An important decision which we took in relation to the way in which the analysis was conducted was to count not only the mistaken forms used by the candidates, but also the correctly used ones, measuring the mistaken forms to them, and obtaining a relevant proportion resulting from the actual ability of the test takers to use vocabulary and grammar in written productions. We used, thus, the method of error analysis described by Corder (1974), but also elements from obligatory occasions analysis (Brown, 1973), usually used in morpheme studies.

The results obtained from counting the forms were introduced in an Excel table, for each category, with the possibilities: correct, mistaken, not leading to misunderstanding and mistaken, leading to misunderstanding, as the last stage of error analysis described by Corder (Corder, 1981). However, with the number of mistakes leading to misunderstanding being rather insignificant in comparison with the number of mistakes not leading to misunderstanding, the two categories were calculated, in the end, together.

We colour coded the types of mistakes (one colour for each type of mistake and for each part of speech), and the correctly used forms (one colour for each category of correctly used forms).

We took the decision of excluding from the corpus the productions which contained more than 50% incorrectly used forms, as they were not representative of the test takers' ability of using elements of grammar and vocabulary correctly at level A1 (3 productions). However, this does not mean that the three test takers did not pass the test of writing; other aspects of their writing, represented by criteria in the grid, like efficiency of communication and complexity, being valued much more in comparison with accuracy.

### 3.3 Results

After all the data were introduced in the Excel tables and calculations made, we decided, according to the results obtained, on how to reflect these results into the new descriptors. The truth is, changes were very significant and the new shape of the grid is substantially different from the old one. The first change concerned grammar. While by grammar we meant in the old grid morphology, we realised that the number of mistakes we discovered in relation to word order, orthography and omitted words was insignificant and that we could very well add them to those of morphology, including them all in the category of grammar. This does not imply that the structure of the descriptors will remain the same for the rest of the levels, the decisions concerning their content being related, at their turns, to the results emerging from corpus analysis. However, for this level, the first very surprising conclusion was that orthography mistakes were in such a small number in the test takers' productions, that a descriptor was not necessary for referring to them. Moreover, in these conditions, the use of it would only confuse the assessor, making him take into consideration a criterion which would indicate a reduced number of mistakes (which would only be normal for the written production) and distracting him from judging other aspects they should consider.

The vocabulary mistakes include the inadequate use of words, and also incorrect spelling. Another significant change in the descriptors concerns the fact that vocabulary and grammar mistakes are now referred to as part of the same descriptor. This decision was taken as a consequence of the fact that, comparing the words with grammar mistakes and the ones with vocabulary mistakes, we found that many of them coincided. Thus, from a percentage of 23.15% words containing grammar mistakes (a percentage which constitutes the cut point between the first two bands), we got to a percentage of 26.13% words containing grammar and/or vocabulary mistakes. By calculating them together we will avoid assessors considering mistaken words twice, once for grammar, once for vocabulary mistakes. The new descriptor concerning grammar and vocabulary for the first band is now formulated like this: "The grammatical and lexical mistakes affect less than a quarter of the words in the text and do not lead to misunderstandings."

While some descriptors were eliminated from our grid or were combined in the same descriptor, new ones were created as a result of corpus analysis. Even some of the best written productions from the corpus contained word transfer, which led us to the decision of creating a new descriptor to reflect this aspect. Thus, with representation of code switching even at the first band of the grid, the assessors will be less tempted to place a production on an inferior band because of instances of word transfer.

The descriptor which, in the old form of the grid, referred to punctuation and orthography was transformed to reflect problems related to punctuation and use of capital letters.

Keeping in mind the suggestions made by the assessors, some other major changes occurred in the structure of the writing grid. Thus, the order of criteria was changed, so as to reflect precedence of efficiency of communication over accuracy. The assessment process was made easier and it was improved through the reduction of the number of descriptors and the introduction of intermediate bands, as well as the assessment of the two tasks separately. The terms used for the creation of the descriptors are much simpler now and are not confusing any longer. The difference between bands is also much clearer and there is no more overlapping between the criteria. Out of the elements important to consider when designing a rating scale, Sara Cushing Weigle identifies as essential two of them: “defining the rating scale, and ensuring that raters use the scale appropriately and consistently” (Cushing Weigle, 2002). The manoeuvrability of the rating instruments contributes immensely to being certain that the grids are used as Cushing Weigle indicates and we were certain that we made a step forward in this sense when, after the presentation of the new shape of the grid, one of our younger assessors commented: “From now on I will not be afraid to rate written productions anymore.”

## 4 Conclusions

Using a corpus of written productions for redesigning the assessment grid proved to be extremely beneficial for our evaluation process from multiple points of view. First of all, the new grid is much clearer and we can count on it for a plus of objectivity, reliability and fairness to the test takers. We can also explain much better, more easily and clearly, for our stakeholders, mainly for our test takers, how the assessment is realised and how points are granted. One of the most interesting results is related to the fact that the information we collected from the processed data based on the corpus provided most of the answers for the observations and suggestions made by the assessors.

This first phase of our project gave us all the reasons to continue with redesigning the writing grids for the levels A2–B2, and also for the speaking grids. This process will provide us not only with a much better assessment instrument, but also with a great source of feedback for our teachers, assessors and test takers, which can only lead to the improvement of the assessment and teaching processes in our department.

## References

Brown, R. (1973). *A first language: The early stages*. Cambridge: Harvard University Press.

Canale M., & Swain M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Corder, S. P. (1974). Error analysis. In J. Allen & S. P. Corder (Eds.) *The Edinburgh Course in Applied Linguistics*, *Volume 3*. Oxford: Oxford University Press.

Corder, S. P. (1981). *Error analysis and interlanguage*. Oxford, UK: Oxford University Press.

Corder, S. P. (1983). A role of the mother tongue. In S. Gass & L. Selinker (Eds.), *Language Transfer in Language Learning* (pp. 85–97). Rowley: Newbury House.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cushing Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Deygers, B., Van Gorp, K., Joos, S., & Luyten, L. (2013). Rating scale design: a comparative study of two analytic rating scales in a task-based test. In E. Galaczi & C. J. Weir (Eds.), *Exploring Language Frameworks. Proceedings of the ALTE Kraków Conference, July 2011* (pp. 271–287). Cambridge: Cambridge University Press.

Ferguson, C. (1971). Absence of copula and the notion of simplicity: a sudy of normal speech, baby talk, foreigner talk and pidgins. In D. Hymes (Ed.), *Pidginization and Creolization of Languages* (219–235). Cambridge: Cambridge University Press.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*. Retrieved from https://www.researchgate.net/publication/249870407

McEnery, T. & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom*. Cambridge: Cambridge University Press.

Platon, E., Sonea, I., & Vîlcu, D. (2012). *Manual de limba română ca limbă străină. A1–A2*. Cluj-Napoca: Editura Casa Cărţii de Ştiinţă.

Platon, E., Sonea, I., Vasiu, L., & Vîlcu, D. (2014). *Descrierea minimală a limbii române. A1, A2, B1, B2.* Cluj-Napoca: Editura Casa Cărții de Știință.

Platon, E. (2016). Two language avatars: The interlanguage and the microlanguage. In I. Boldea (Ed.), *Globalization and National Identity. Studies on the Strategies of Intercultural Dialogue* (pp. 634–647) . Târgu Mureş: Arhipelag XXI.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics, 10*, 209–231.

# The Evaluation of Chinese Students of Italian L2 in China and in Italy: Practices at the Universities for Foreigners of Perugia and Siena

**Giuliana Grego Bolli**, University for Foreigners of Perugia, Italy
**Sabrina Machetti**, University for Foreigners of Siena, Italy
**Danilo Rini**, University for Foreigners of Perugia, Italy
**Anna Bandini**, University for Foreigners of Siena, Italy

**Abstract:** The assessment of Chinese students' proficiency in Italian is an issue concerning a lot of institutions, be they universities or not, in order to let the students enroll in undergraduate or high specialization courses. It is a part of the tensions which have characterized the scientific debate on Language Testing in relation to the positive and negative consequences of language testing, and in particular of the standardized assessment on Italian L2 teaching and learning, particularly in relation to Chinese students. As a result, the institutions dealing with language certification need to provide evidence of the exam's evaluation process, to guarantee that the results and their consequent use are valid and reliable. From this perspective the paper provides a comparison between the results obtained in two certification exams provided by the University for Foreigners of Perugia, and of Siena, in China and those obtained by the same students in Italy after an eight-month period, enhancing strengths and weaknesses.

## 1 Introduction

The aim of this contribution is to look at the performance of groups of Chinese students who want to enter the University in Italy or to study Music or Arts taking the CELI examinations, awarded by the University for Foreigners of Perugia, or the CILS examinations awarded by the University for Foreigners of Siena. Points of strength and weakness in the performance in the CELI or CILS examinations of two different relatively small groups of students will be highlighted.

All the students were involved in a specific joint project between China and Italy called Progetto Marco Polo Turandot (MPT Project). Thanks to this project started in 2007, the number of university students as well as exchange students from China to Italy has sharply risen in the past few years.

The MPT students generally start to study Italian for three or four months in China, continuing their language tuition in Italy for around eight months. In order to enter Italian Universities or Academies of Art or Conservatories of Music, MPT students need to have a B2 level language certificate; sometimes also a B1 level certificate can be accepted.

As Scibetta (2016) discusses, from a social and cultural point of view this phenomenon has considerably contributed to changing some diffused social representations and negative social stereotypes systematically attributed to the Chinese population by the hosting society and by Italian media. From an educational point of view, this phenomenon has likewise had a strong impact (Diadori & Di Toro 2009; Rastelli, 2010; 2013) especially referred to learning and teaching Italian as an FL and teaching and assessment materials. Chinese students, as any other international student, need to be competent in certain language areas and skills to be able to cope with academic demands (Cumming, 1994). Despite not being among the purposes of this contribution, this issue needs to be addressed systematically, the 'language areas' need to be described and what exactly 'to be able to cope with academic demands' implies needs to be defined in order to validate the overall assessment process.

The preliminary results of the analysis run on the performance on CELI and CILS examinations could offer a very initial empirical perspective, data-driven, to reflect on a more systematic and shared training into Italian of Chinese students in terms of needs, objectives, contents and duration. For the purpose of this contribution, it is also important to highlight that the educational assessment has a very strong impact within Chinese society. Examinations still play an important social and educational role in China as fair measurement for selection into the social hierarchy (Cheng & Qi, 2006). Cheng (2008) traces back to 2,000 years ago the tradition of formal testing and assessment within the Chinese educational system, hence Chinese students are very well used to formal assessment, fully aware of its impact and consequences in their University and professional career.

## 2 CELI Exams: structure

All the CELI examinations test all areas of language ability (Grego Bolli & Spiti, 2004). They are divided into two parts: the written part and the oral one, and made up of different papers: from three in CELI Impatto (A1) to five in CELI 3 (B2). The Speaking test is a face to face structured interview with one candidate, an interlocutor and an examiner. The written part of the examinations is centrally marked (in Perugia) by very experienced examiners. The Speaking test is marked locally (in the Examination Centres) after specific training sessions run by CVCL staff both in Perugia and locally. The certificate is awarded only if candidates pass both the written and the oral part of the exam. If a candidate passes only one part, s/he can keep this result for one year. Five CELI examination levels (from A2 to C2) have been awarded the ALTE Q-mark after passing an audit, and meeting the 17 ALTE quality standards.

## 3 Results achieved by candidates under scrutiny in CELI exams

This contribution takes into consideration the results achieved by two groups of Chinese students of the Nanjing Normal University after taking CELI exams both in China and in Italy. Though the numbers of students involved are small, the results they achieved show some tendencies that may be of help for further data collection, and may also give interesting hints on how the linguistic education the students received, as well as their assessment, may be better tailored to the needs of this specific group of language learners.

The first group under scrutiny was made up of 42 students (N=42), and their results were monitored during 2014 and 2015. In November 2014, after four months of education in Italian language in Nanjing, 18 of them took the CELI Impatto exam (that is A1 level), and the rest (n=24) took CELI 1 (A2). After that, they moved to Perugia, where they studied Italian at the University for Foreigners for eight months and finally took another CELI exam at a higher level in August 2015. In detail, 30 out of 42 took CELI 2 (B1) and only 12 took CELI 3 (B2).

Given the small number of candidates, for the sake of clarity, and also to maintain a closer comparability with CILS exams, the results here are grouped into level bands A and B, rather than in single CEFR levels. In Nanjing for A band exams (*CELI Impatto* and CELI 1) the results show that 28 out of 42 students passed the whole exam (over 66%), whereas only 4 of them failed. Of the remaining students, 10 (around 23%) passed one part only; in detail, 4 passed the oral, and 6 the written part.

These figures did not match the results in Perugia, where only 28% (12 out of 42 students) passed the CELI B band exams (CELI 2 and CELI 3), and 26% failed. 45% passed one part only (n=19; 17 of them passed the oral part, whereas only 2 passed the written part). It has also to be noted how 30 students took the B1 level exam and only 12 took the B2.

By comparing the results achieved in the different papers of the two sessions, it turns out also that, after around eight months of tuition in Italy, candidates scored very low on the receptive skills compared to the productive ones, and more generally it seems very difficult to consolidate, to reinforce the passage from A to B band levels. Some explanation of this outcome may include: students being more trained on some abilities rather than others; guessing answers to objective items in receptive skills; quite lenient examiners and markers. Cheng and Gao (2002, p. 22–23) showed that guessing was prevalently used by a sample of students taking the multiple choice reading comprehension test of the College English Test (CET) in China.

Moving to the next group in 2015–2016, even though with smaller numbers of students, some similar tendencies may be noticed.

The second group under scrutiny was made up of 23 students who took CELI 1 (A2) in November 2015. 11 passed the whole exam (around 48%), 8 failed (35%), and 4 (17%) passed one part only (3 the oral and 1 the written part). In Perugia, taking also into account the poor results achieved in 2015, CELI 2 and 3, a specific exam tailored to the tuition received by students in MPT courses, was produced. This resulted in a better performance of the group of learners, with 15 students out of 23 (over 65%) passing all B band exams (16 students took CELI 2 MPT – B1, whereas 7 took CELI 3 MPT – B2); a mere 2 out of 23 failed (below 9%), and 6 out of 23 (26%) passed the oral part only.

By comparing the results, it seems quite evident that the second group of candidates performed better than the first group, the previous year, in Perugia. Also the percentage of candidates who reached the cut-off scores in each paper, after around eight months of tuition in Italy, is higher in comparison to the first group. The weakest performance is in both cases in listening comprehension, even though in 2016 students reaching the cut-off score for this skill were over 20 percentage points above 2015 (47% vs. 26%).

In conclusion, despite the small sample of students under scrutiny, Chinese candidates seem to perform better in the CELI MPT exams, than they do in the general purpose CELI 2 and 3.

### 4 CILS exams: structure

As with CELI, all the CILS examinations test all areas of language ability. They are divided into two parts: the written part (Listening, Reading, Grammar, Writing) and the oral one (Speaking). The Speaking test consists of two tasks: a face to face structured interview with one candidate and an interlocutor; a monologue. This test and the tests from the written part are centrally marked (using an assessment online platform – Bandini, Lucarelli, Sprugnoli & Strambi 2012) by continuously trained raters. The certificate is awarded only if candidates pass both the written and the oral part of the exam. If a candidate passes only one part, s/he can keep this result for 18 months (Machetti, 2016; Vedovelli, 2005).

## 5 Results achieved by candidates under scrutiny in CILS exams

This contribution takes into consideration the results achieved by two groups of Chinese students after taking CILS exams both in China and in Italy: the first one of the Hebei Normal University, the second one of different Chinese universities in Beijing, Chongqing, Shanghai, Wuhan. As in the CELI exams, though the numbers of students involved are small, the results they achieved show some tendencies that may be of help for further data collection, and may also give interesting hints on how the linguistic education the students received, as well as their assessment, may be better tailored to the needs of this specific group of students.

The first group under scrutiny was made up of 61 students in China and 59 in Italy, and their results were monitored during 2014 and 2015. In December 2014, after four months of education in Italian language in Hebei, 21 of them took the CILS A2 exam, 40 took CILS UNO – B1. After that, they moved to Siena, where they continued to learning Italian at the University for Foreigners for eight months and finally took the MPT CILS exam (a specific B1+/B2- exam, tailored to the tuition received) in August 2015.

In Hebei for CILS A2 exams the results show that 12 out of 21 students passed the whole exam (over 57%), 9 passed one part only, and nobody failed; for CILS UNO – B1 exams the results show that 11 out of 40 students passed the whole exam (over 27%), 25 passed one part only, only 2 failed, and 2 candidates didn’t attend the exam.

These figures did not match completely the results in Siena: for the candidates that in China took CILS A2 exams, the results show that 9 out of 20 students passed the whole exam (over 47%), 6 passed one part only, 4 failed, and 4 candidates didn’t attend the exam. For the candidates that took CILS UNO – B1 exams in China, the results show that 26 out of 39 students passed the whole exam (over 66%), 4 passed one part only, only 1 failed, and 8 candidates didn’t attend the exam.

By comparing the results achieved in the different papers of the two sessions, it turns out also that, after around eight months of tuition in Siena, candidates that in China took CILS A2 exams scored lower on the Listening, Reading and Grammar, and more in general it seems very difficult to reinforce the passage from A2 to B1/B2 levels. Referring to candidates that took CILS UNO – B1 exams in China, they scored better on the Listening, Reading and Grammar, although it seems very difficult to speak of a B2 proficiency in Italian. Some explanation of this outcome may include students being more trained on some abilities rather than others, lenient examiners, etc.

The second group under scrutiny in 2015–2016 was made up of 80 students, who took CILS A2 and CILS UNO – B1 in December 2015. For CILS A2 exams the results show that only

4 out of 46 candidates passed the whole exam (over 8%), 42 passed one part only, and nobody failed. For CILS UNO – B1 exams, the results show that only 3 out of 34 candidates passed the whole exam (over 8%), 28 passed one part only, and only 3 out of 34 failed. Listening and Speaking were the most difficult parts for the CILS A2 candidates; Listening, Reading and Grammar for the UNO – B1 candidates. In Siena (August 2016), 47 students took the CILS UNO – B1 exam and 15 the CILS DUE – B2 exams. For CILS UNO – B1 exams the results show that 20 out of 47 candidates passed the whole exam (over 42%), 27 failed, and nobody passed one part only. For CILS UNO – B1 exams, the results show that 10 out of 15 students passed the whole exam (over 66%), 5 failed, and nobody passed one part only. None of the candidates showed significant differences among the parts of the exam, also if Listening, Reading and Writing were more difficult then Grammar and Speaking.

## 6 Conclusion

The first results of the analysis run on the performance on CELI and CILS examinations could offer a very initial empirical perspective to reflect on a more systematic and shared training in Italian of Chinese students in terms of needs, objectives, contents and duration. Certainly, MPT students need to be more trained in Italian, in term of duration and contents. Some abilities need to be more trained than others: for example, Scibetta (2016) focuses on Chinese University students' acquisition of L2 Italian from the point of view of pragmatics and asks to deepen the knowledge about the difficulties as well as the specificity of this category of students in developing particular pragmatic skills in Italian. Secondly, he suggests some effective teaching techniques, in order to help Chinese students develop their pragmatic competence faster, which could be also adapted to other categories of language learners.

Our analysis highlights the need to reflect also on the phenomenon of guessing answers to objective items in receptive skills and the impact of this behaviour on the examinations results. Referring also to Chinese society, Fulcher (2011) suggests that "tests are a means of introducing meritocracy to society, and for them to work it is essential that they are valued in their own right as a means to establish and maintain fairness in society. The 'people thing' must be about educating the general population to understand the role of tests in society and why cheating undermines the principles of liberal meritocratic states."

## References

Bandini, A., Lucarelli, S., Sprugnoli, L., & Strambi, B. (2012). Procedure di verifica della valutazione nei test di certificazione. In S. Cacchiani, S. Morgan & M. Silver, *Standardized Language Testing: Contemporary Issues and Applications, Rassegna Italiana di Linguistica Applicata*. Roma: Bulzoni.

Cheng, L., & Gao, L. (2002). Passage dependence in standardized reading comprehension: Exploring the College English Test. *Asian Journal of English Language Teaching, 12*, 161–178.

Cheng, L. & Qi, L. (2006). Description and examination of National Matriculation English Test in China. *Language Assessment Quarterly, 3*(1), 53–70.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*, 15–37.

Cumming, A. (1994). Alternatives in TESOL research: Descriptive, interpretive, and ideological orientations. *TESOL Quarterly, 28*, 673–703.

Diadori, P., & Di Toro, A. (2009). Come insegnare italiano agli studenti di madrelingua cinese? Un'introduzione. In P. Diadori (a cura di), *La DITALS risponde 6*. Perugia: Edizioni Guerra, 66-77.

Fulcher, G. (2011). *Cheating on Language Tests*. Retrieved from http://languagetesting.info/features/examination/cheating.html

Grego Bolli, G., & Spiti, M. G. (2004). *Misurare e valutare nella certificazione CELI*.

Machetti, S. (2016). Verificare, misurare, valutare l'italiano di stranieri. Il caso della CILS. *Revista de Italianistica, 32*, 94–114.

Rastelli, S., a cura di (2010). *Italiano di cinesi, italiano per cinesi*. Perugia: Edizioni Guerra.

Rastelli, S., a cura di (2013). *Il processing nella Second Language. Teorie, dati sperimentali, didattica*. Roma: Carocci.

Scibetta, A. (2016). Chinese University Students' Development of Pragmatics Skills in L2 Italian. A Corpus-Based Study. In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics, 4*, 243–271.

Vedovelli, M., a cura di, (2005). *Manuale della certificazione dell'italiano L2*. Roma: Carocci.

# Verifica delle possibilità per una certificazione della lingua araba

**Aisha Nasimi**, Università per Stranieri di Siena, Italia

**Abstract:** La ricerca mira alla verifica delle possibilità per una certificazione della lingua araba, come lingua straniera, e allo sviluppo di un proto-test sperimentale basato su uno specifico profilo di apprendente in un preciso contesto di acquisizione, strutturato sugli attuali test delle lingue europee, verificando l'applicabilità degli standard e i principi del Quadro comune europeo di riferimento (QCER), a partire da un'analisi storico-linguistica e sociolinguistica e da indagini mirate alla definizione delle motivazioni dell'apprendimento di questa lingua. La valutazione della lingua araba rappresenta un complesso oggetto di ricerca: il fenomeno della "diglossia" (Ferguson, 1959) e la dimensione sacrale e di prestigio della lingua giocano un ruolo primario nell'analisi sociolinguistica di questo idioma, allo stesso tempo, oggi l'arabo è considerato una delle 'lingue emergenti' all'interno del cosiddetto "mercato globale delle lingue" (Calvet, 2002). Tuttavia, la marcata variazione linguistica fra 'norma' e 'uso' rappresenta una notevole sfida per stabilire uno standard linguistico uniforme e condiviso, utilizzabile in un test linguistico, che possa assicurare il massimo grado di spendibilità sociale, obiettivo principale di una certificazione, in un mondo dove la crescente mobilità delle persone e il conseguente rilevante contatto linguistico incoraggiano la formazione di nuove identità.

## 1 Introduzione

La ricerca ha come obiettivo la verifica delle possibilità per una certificazione della lingua araba, come lingua straniera in Italia, con lo sviluppo di un test pilota sperimentale basato su specifici profili di apprendenti in precisi contesti di acquisizione, strutturato sul modello degli attuali test delle lingue europee, verificando l'applicabilità degli standard e dei principi del QCER alla lingua araba, a partire da un'analisi di tipo storico linguistico e sociolinguistico.

## 2 Motivazioni

Le motivazioni che stanno alla base della ricerca sono connesse al fatto che la lingua araba sia di fatto oggi una delle 'lingue emergenti' all'interno del "mercato globale delle lingue" (Calvet, 2002; De Mauro, Vedovelli, Barni, & Miraglia, 2002), che sia inoltre una delle lingue più rilevanti al mondo per numero di parlanti (la quinta secondo i dati Ethnologue). Oltre ad essere la lingua ufficiale di 22 paesi e una delle sei lingue ufficiali delle Nazioni Unite, essa costituisce, nella sua versione standard, la lingua franca per l'intercomprensione degli arabofoni in tutto il mondo, così come il simbolo dei tratti condivisi della cultura arabo-islamica che attraversano l'area araba in maniera trasversale. Oltre ad essere presente ormai da anni in Europa come lingua di immigrazione, si registra a partire dai primi anni 2000 un aumento degli apprendenti e quindi dell'offerta formativa che prevede l'insegnamento della lingua araba sia in contesto europeo che in quello statunitense. Non essendo quindi ancora stata sviluppata una certificazione per la lingua araba che misuri le quattro abilità su larga scala sulla base di criteri e standard condivisi in Italia, è quanto mai necessario oggi poter predisporre di tale strumento come garanzia per la spendibilità sociale di questa lingua, così come avviene per le altre lingue.

## 2 Metodo

Il metodo della ricerca si sviluppa a partire da un'analisi di tipo sociolinguistico e storico-linguistico della lingua araba, con lo scopo di delineare un possibile modello di lingua di riferimento per un test di certificazione; in una fase successiva, sono state poi messe a punto due indagini a campione per la definizione dei profili degli apprendenti, in questo specifico caso rivolte ad apprendenti adulti, tramite due questionari, di cui uno diffuso fra studenti universitari e uno a cui hanno risposto studenti di lingua araba adulti appartenenti ad un pubblico misto. L'ultima fase della ricerca, attualmente in corso, unitamente ad una mappatura dettagliata delle certificazioni di lingua araba ad oggi esistenti nel mondo e ad un'analisi delle questioni generali intorno alla valutazione linguistica di questa lingua, prevede lo sviluppo di un test pilota sperimentale di lingua araba come lingua straniera somministrato a due campioni di studenti universitari, uno di livello comparabile al livello A1 del QCER e l'altro di livello A2, a cui segue una validazione tramite questionario di valutazione per i candidati e l'analisi degli item che costituiscono i test.

### 3 Le fasi della ricerca

La sezione dedicata all'analisi sociolinguistica della lingua araba ha previsto un'approfondita descrizione degli aspetti salienti che caratterizzano questa lingua che interessano gli usi e la variazione linguistica, ma anche la dimensione di prestigio, i quali, in questo caso specifico, sono strettamente connessi. Il fenomeno della "diglossia" (Ferguson, 1959) che caratterizza questa lingua prevede infatti una forte dicotomia, ma allo stesso tempo la coesistenza in un continuum da una parte della lingua standard unitamente alla lingua classica come lingue di prestigio legate sia alla dimensione sacrale (connessa con il testo coranico) che al corpus letterario antico canonico, i quali rispecchiano la norma, la lingua "corretta", e dall'altra i dialetti, le lingue della quotidianità e dei contesti più informali, simbolo della variazione linguistica e della dimensione d'uso. Il modello di lingua più consono ad oggi per un test di lingua araba è costituito dall' MSA (Modern Standard Arabic), nato sulla scia della "riforma linguistica" di carattere per lo più lessicale a partire dal XIX secolo ad opera delle Accademie della Lingua Araba con l'obiettivo di preservare ma allo stesso tempo di sviluppare la lingua, integrandola con un lessico rinnovato e al passo con i tempi moderni. L'MSA è stato di fatto al centro delle politiche linguistiche dei paesi arabi a partire dalla loro indipendenza ed è oggi in via di semplificazione per l'influenza delle lingue europee (un esempio è il cambiamento dell'ordine dei costituenti della frase da VSO a SVO). Simbolo identitario condiviso di "arabicità" del patrimonio storico-culturale comune del mondo arabo, in quanto direttamente discendente dall'arabo classico, il MSA è il modello linguistico che si rifà alla norma codificata condivisa e riconosciuta e dal quale gli arabofoni attingono sia per la comunicazione formale che per la comprensione reciproca, con lo scopo di poter quindi superare le differenze che caratterizzano i propri dialetti di provenienza (Bassiouney, 2009; Durand, 2009). Tuttavia, pur non essendo di fatto l'MSA la lingua nativa di nessun arabofono, bensì una lingua acquisita pienamente con la scolarizzazione, l'esposizione ad essa inizia molto precocemente, cosa che permette loro di sviluppare presto una competenza almeno di tipo passivo (Albirini, 2016). I dialetti dall'altra parte, anche detti neoarabi, molto spesso ricollegati ad una versione corrotta e sgrammaticata della lingua, sono legati per lo più ai contesti d'uso della lingua più informali, pur operando una forte spinta decentratrice sulla lingua standard (che comprende l'arabo classico e l'MSA). La complessità della situazione sociolinguistica del mondo arabo rende perciò alquanto ardua l'individuazione di un costrutto autentico valido per un test linguistico, comportando il rischio di una mancata corrispondenza con la reale competenza comunicativa del parlante nativo. Questo è, infatti, uno degli aspetti più critici per la valutazione della lingua araba, il quale rende senza dubbio la costruzione di un test affidabile una sfida a tutti gli effetti.

La seconda fase della ricerca, relativa alla definizione del profilo dell'apprendente adulto di lingua araba, pone come questione centrale l'individuazione delle motivazioni principali allo studio di questa lingua. L'indagine, che al momento risulta essere la prima in Italia condotta su tali tematiche, si è articolata nella stesura e nella diffusione di due questionari a cui hanno risposto 174 studenti universitari e 121 studenti adulti appartenenti ad un pubblico misto. I questionari, attualmente in fase di analisi dei risultati, evidenziano come le principali motivazioni allo studio della lingua araba che interessano il pubblico adulto siano legate nella maggior parte dei casi ad un forte interesse e curiosità verso questa lingua e alla relativa cultura, ma anche per favorire le dinamiche interculturali e quindi il contatto fra lingue e culture, unitamente alla motivazione strumentale connessa alla possibilità di avere maggiori opportunità nel mondo del lavoro. La percezione della lingua araba da parte di chi si è approcciato al suo apprendimento è prevalentemente positiva, è inoltre considerata come lingua rilevante a livello globale, anche se permane l'idea di una complessità che la caratterizza, riflettendo quindi l'immaginario ad essa abitualmente connesso. Nello specifico, le abilità ritenute più difficili da acquisire sono quelle del parlato e dell'ascolto, dato che evidenzia come il fenomeno della diglossia si ripercuota in qualche modo anche nel processo di apprendimento dell'arabo. Le indagini evidenziano pertanto come lo studio dell'arabo possa essere considerato come valore aggiunto in ambito lavorativo, oltre ad essere lo strumento chiave per una maggiore comprensione degli aspetti geopolitici che interessano l'attualità dei giorni nostri, e soprattutto per favorire le dinamiche interculturali, costituendo quindi lo strumento decisivo dal quale partire e del quale munirsi per poter superare stereotipi e luoghi comuni oggi tipicamente legati al mondo arabo-islamico.

La terza e ultima fase della ricerca, attualmente in corso, è legata più strettamente al tema della valutazione linguistica della lingua araba oggi in Italia. Essa si articola a partire da una dettagliata mappatura dei test di certificazione di lingua araba attualmente esistenti nel mondo, con lo scopo di avere un quadro d'insieme delle caratteristiche e dell'avanzamento del campo dei test di lingua araba oggi. A ciò segue la presa in esame del settore della valutazione della lingua araba in genere, del suo esordio e del suo sviluppo in contesto statunitense dove sono stati messi a punto i primi test standardizzati di competenza di tipo proficiency di arabo, oltre che ai test linguistici di tipo diagnostico e di tipo achievement, a partire dagli anni '60-'70 (Rammuny, 1999). Sono state poi oggetto di analisi le principali criticità che interessano in maniera specifica il testing di una lingua diglossica come l'arabo, proponendo dei modelli interpretativi e possibili approcci al riguardo. Segue una riflessione approfondita circa la possibilità di applicare gli standard e i criteri del QCER alla valutazione della lingua araba. In quanto lingua non europea, sono infatti molti i dubbi e le perplessità oggi da parte di molti sulla effettiva possibilità di impiego di uno strumento pensato appositamente per le lingue europee, anche a lingue tipologicamente diverse come l'arabo. Tali aspetti vengono poi confrontati con quanto portato avanti dall'ILR negli Stati Uniti, con le linee guida ACTFL (ACTFL, 2012) e le annotazioni specifiche per la lingua araba, fornite a partire dal 2012, in cui si evidenzia una maggiore attenzione alle specificità e alle caratteristiche tipologiche di questa lingua.

L’ultima fase della raccolta dei dati è consistita nella costruzione e somministrazione di due test pilota sperimentali per la misurazione delle competenze generali della lingua araba su due campioni di studenti universitari, di livello corrispondente al livello A1 e A2 del QCER, i quali quindi hanno frequentato rispettivamente un’annualità e due annualità di studio della lingua in questione. I test, strutturati in maniera analoga ai test di competenza per le lingue europee, prevedono quesiti per la misurazione delle quattro abilità, quali ascolto, comprensione della lettura, scrittura e parlato, distribuite in sette prove di cui due di ascolto, due di comprensione della lettura, due di grammatica e strutture della comunicazione e una prova scritta a cui  segue la prova orale con interazione faccia a faccia somministrata ad un piccolo gruppo di studenti fra coloro i quali hanno svolto il test. In una fase antecedente al test è stato messo a punto un sillabo specifico per la lingua araba contenente le specificazioni dei test, quindi l’insieme degli elementi e delle strutture linguistiche che possono potenzialmente essere inclusi al suo interno sia per il livello A1 che per il livello A2, unitamente ad un sillabo delle tipologie testuali relativo ai profili degli apprendenti e dei contesti d’uso attinenti ad ogni prova. La fase della validazione dei test si sviluppa a partire dall’assegnazione di un punteggio finale, secondo criteri di valutazione precisi specificati in un’apposita griglia di valutazione che prevede un punteggio di uguale peso attribuito alle differenti prove. Ad essa segue un’analisi dei dati acquisiti di tipo qualitativo tramite un breve questionario di valutazione da parte degli studenti che hanno partecipato alla somministrazione, in cui viene chiesto di attribuire un giudizio per ciò che concerne la difficoltà di svolgimento dei quesiti a seconda delle differenti abilità testate. Successivamente segue l’analisi dell’adeguatezza degli item, nello specifico del loro indice di difficoltà, tramite il calcolo degli indici di Item Facility (IF), con lo scopo di individuare eventuali aspetti problematici legati alla costruzione dei quesiti e alla scelta dei testi rispetto ai risultati ottenuti nel test e quindi alle capacità e competenze in lingua araba proprie degli studenti.

## 4 Conclusioni

I risultati attesi dalla ricerca includono la possibilità di poter disporre presto dello strumento certificatorio per le quattro abilità anche per la lingua araba in contesto italiano, basata su criteri e standard di valutazione condivisi, come documento che verifichi le competenze linguistiche di chi ha l’esigenza di attestare la propria conoscenza della lingua araba per scopi professionali o legati alla propria formazione. Si auspica inoltre che, essendo ad oggi tale ambito di ricerca oggetto di un numero piuttosto scarso di studi, esso possa essere ampliato, soprattutto per quanto riguarda la raccolta di dati più approfonditi concernenti la validità e l’affidabilità dei test. Da una prima analisi dei dati raccolti con la presente ricerca emerge inoltre come, oggi ci sia un forte bisogno di aggiornamento delle tecniche glottodidattiche della lingua araba, in un’ottica di apprendimento ‘integrato’, tenendo quindi conto anche degli usi autentici della lingua: solo così è possibile soddisfare le esigenze dei nuovi apprendenti, oggi sempre più proiettate verso l’apprendimento della lingua dell’uso e quindi all’acquisizione di una competenza linguistico-comunicativa effettiva.

Ciò che è più auspicabile quindi, è una maggiore attenzione e incoraggiamento allo studio della lingua araba, e che questa tendenza si possa tradurre sempre più nel tempo nella promozione di questa lingua allo scopo di favorire il contatto fra lingue e culture in chiave plurilingue e di conseguenza il dialogo all’interno delle società odierne.

## Letture di approfondimento

Angelescu N. (1993). *Linguaggio e cultura nella civiltà araba*. Torino: Silvio Zamorani Editore

Consiglio d’Europa. (2002). *Quadro Comune Europeo di Riferimento per le lingue: apprendimento, insegnamento, valutazione*. Milano: La Nuova Italia.

Davies A. (1990). *Principles of Language Testing*. Oxford: Blackwell.

Mc Namara T. (2014). *Language Testing*. Oxford: Oxford University Press

Veersteegh K. (2014). *The Arabic Language*. Edinburgh: Edinburgh University Press.

Wahba M. K., & Taha Z. A., & England L. (Eds.). (2006). *Handbook for Arabic Language Teaching Professionals in the 21st Century*. Mahwah: New Jersey

## Bibliografia

ACTFL. (2012). *ACTFL Proficiency Guidelines 2012 – Arabic; Annotations and Samples* Retrieved from httlp://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/arabic

Albirini A. (2016). *Modern Arabic Sociolinguistics. Diglossia, variation, codeswitching, attitudes and identity*. Abingdon; New York: Routledge.

Bassiouney R. (2009). *Arabic Sociolinguistics. Topics in Diglossia, Gender, Identity, and Politics*. Edinburgh: Edinburgh University Press.

Calvet L. J. (2002). *Le marché aux langues. Les effets linguistiques de la mondialisation*. Paris: Les Edition Plon.

Durand O. (2009). *Dialettologia araba*. Roma: Carocci.

Ethnologue https://www.ethnologue.com/

De Mauro T., Vedovelli M., Barni M., & Miraglia L. (2002), *Italiano 2000. Indagine sulle motivazioni e sui pubblici dell’italiano diffuso fra stranieri*. Roma: Bulzoni.

Ferguson, C. A. (1959). La diglossia. In P. Giglioli & G. Fele (Eds.) (2000), *Linguaggio e contesto sociale* (pp. 185–205). Bologna: Il Mulino.

Rammuny, R. M.(1999). Arabic Language Testing. *Al-‘Arabiyya*, *32*, 161–194.

# The Implementation of a French Language Certification: Positive Washback and Wider Resulting Effects

**Stéphanie McGaw**, University of Corsica, France

**Abstract:** Since the generalising of language teaching in French universities as an answer to the European-driven will to improve the students' language level, language assessment has become an important issue and aroused questions which had hardly ever been crucial before. This paper shows how the implementation of a French multilingual certification, the CLES (Higher Education Language Certification), was a positive answer to issues at micro (individuals), meso (learning) and macro levels (institution and educational system) and appears as a lever of change and good practice in the French higher education system. In order to demonstrate the positive impact of this certification, we will proceed to a washback study. Our approach being that of complexity, we will then propose a systemic analysis of the CLES organisation. This will allow us to answer the Bailey model (1996, 2017) and propose a model of positive washback after Bailey's.

## 1 Introduction

Since the generalising of language teaching after the LMD (Licence-Master-Doctorate) reform, which underlined the necessity to improve the students' language level and allow their mobility, language assessment has become an issue in French universities and aroused questions which had hardly ever been crucial before.

The aim of this paper is to show how the implementation of the CLES (Higher Education Language Certification) at the University of Corsica was a positive answer to issues at micro (individuals), meso (learning) and macro levels (institution and educational system). The CLES thus appears as a lever of change and good practice in the French higher education system.

In order to demonstrate the positive impact of this certification, we will proceed to a washback study. Our approach being that of complexity, we will then propose a systemic analysis of the CLES organisation. This will allow us to answer the Bailey model (1996, 2017) and propose a transferable model of positive washback.

## 2 Impacts of the Implementation of the CLES

With the LMD reform (2003), which finds its roots in the Bologna process (1999), universities were endowed with the new mission of improving the students' language competencies in order to allow and increase their mobility. The European-driven will, translated in a language-for-all policy, logically aroused the question of assessment.

### 2.1 Issues to be Addressed

Poteaux (2014) underlines the situation in which University Presidents found themselves and gives a list of the questions they had to find answers to:

Each university institution decides on its foreign language policy. Languages, what for? General culture or professional project? Obligatory or optional? Included in the 30 semester credits Units or in addition, out of the curriculum? Subject to compensation between teaching units, independent, or self-sufficient? Are they solely the decision of each faculty or are they part of a general policy that is reflected in all the training curriculum? In close connection with specialty subjects or general language in extension of secondary education? English predominant or diversified offer? Which certifications to choose or impose on students, and who finances them? (Poteaux, 2014, p. 4)

In addition to this situation, one must bear in mind that satisfying the reform implied a deep change in the philosophy of what teaching meant at the university. Fave-Bonnet (1994) highlights the difficulty for university teachers to change their practice and posture, since their role, as stated in the decree 84-431 (June 6th, 1984) is that of “insur[ing] the transmission of knowledge”. She underlines the fact that “they are trained for research (and usually not for teaching), they are recruited on scientific criteria (and not on pedagogical criteria), and they are often honoured for their research renown rather than for their teaching popularity” (Fave-Bonnet, 1994, p. 13).

Indeed, as teachers in French universities do not benefit from any training, it is quite difficult to think they can evolve towards an action-orientated approach and think their mission and role differently. What has to be retained from the context is the difficulty to satisfy to a reform nobody was prepared for, and the difficulty to input cross-disciplinarily in a world of specialists.

### 2.2 Brief Presentation of the CLES

This certification was created in May 2000 by LANSAD (language for specialists of other disciplines) teachers in order to promote multilingualism and trigger an evolution in teaching practices. It is accredited by the Ministry of Higher Education and Research. This scenario-based certification is based on the Common European Framework of Reference for Languages (CEFR) and available in nine languages (English, German, Spanish, Portuguese, Italian, Arabic, Polish, modern Greek and Russian), at three CEFR levels (from B1 to C1). The subjects are linked to the students’ ‘interests, such as mobility, job training or society issues, which allow them to put into practice the four skills of communication, to which must be added the interaction skill. There are around 25 000 candidates a year who sit this exam in one of the 54 centres belonging to the 11 regional hubs of the territory. The national coordination (University of Grenoble) is composed of a scientific director and four assistant directors (general affairs, subjects B2–B1/C1 and training), an international scientific comity, and a piloting committee.

In order to implement the CLES each university has to be accredited by the Ministry. The CLES functions on the principle of a pooling system: each university member contributing to the national organisation by depositing exam subjects on the national platform, which will be dispatched in the centres nation-wide. Each validation by the national coordination opens a credit of five exam subjects for each contributing university. Certification writers must primarily go through a training session insured by the national coordination, before they are allowed to send exam subjects.

### 2.3 What the Implementation of the CLES Changed in Our University: Washback Study

Here we will give a description of the effects observed in situ after the implementation of the CLES, and confirmed at a national scale via a Google Form enquiry. We will first focus on the effects observed at the micro level, since they are the more visible ones and will then consider the impacts at meso and macro levels, which we consider as the consequences of micro effects. Bachman & Palmer (1996) develop the definition of washback given by Alderson & Wall (1993), who see washback as the effects of a test at a micro level, on teaching and learning. Bachman & Palmer (1996, p. 29–30) consider washback as being one dimension of impact, which they define as the effects a test can have on an educational system or society as a whole.

### *2.3.1 Washback*

Before the LMD reform and the implementation of the CLES, language teachers were in charge of a non-fundamental subject within curricula of other specialities than languages, and most of them were temporary workers entrusted with the teaching of some (usually) English lessons. With the implementation of the CLES, the need to recruit dedicated teachers arose.

The first impact was then on the teacher, who from that moment belonged to a local team thanks to the definition of a transversal project and the definition of a common objective, namely that of certifying as many students as possible at a B2 level. The certificate format triggered a pedagogical reflexion, and the fact that it was built on to the CEFR made the teachers access the oriented-action paradigm. Moreover, the fact that they should take part in the pooling system by writing subjects gave them an active role to play at a national level, being thus part of a national team, gaining expertise and recognition. We can define this impact as both social and academic since joining the CLES added value to the mission of the LANSAD teacher, but also to the quality of the training offered by the institution. By highlighting the usefulness of the CLES, we are here following the Bachman and Palmer (1996, p. 17) statement as regards the quality of a test: “the most important quality of a test is its usefulness”. Our study, based on class observations and questionnaires, proved the CLES is also useful to the students for two main reasons: they attend lessons (which was not the case before). They now play an active role in their training via actual role-playing situations, and get transferable competencies which makes them aware of the economic context in which they will evolve. The CLES allows them to get official proof of their language competencies, which makes them more employable

### *2.3.2 Impacts*

The objective for the student to get a CLES B2 level, certifying that he is able to do specific tasks through language use, triggered a harmonisation of teaching modes at an institutional level. The contents, which were formally defined by each (non-permanent) teacher, and the training offer were then equal and extended with the multiplication of language choice. At the institutional level, the CLES allowed a gain in quality with the definition of a language policy: publication of specific teaching positions, choice of languages and CEFR level. The pedagogical evolution triggered by the CLES format and functioning helped design a global project which opened on the building of an international centre integrating a language and certification centre as well as the international relations service.

### *2.3.3 Aftereffects*

We identify a third type of effect we define as “aftereffects”, a sort of de facto effect. We define them as consequential effects of washback and impact effects. They are gains for both the individuals (empowered) and the institution.

The most important aftereffects concern meso and macro levels with harmonised and Europeanised curricula which make the academic offer more readable and the institution more visible within the European Higher Education Area (EHEA), jointly promoted by the Bologna process (1999) and the Lisbon process (2000). At a national level, and within the context of the assessment by the High Council for the Evaluation of Research and Higher Education (HCERES), the institution can demonstrate it has an effective language policy meeting European recommendations.

The effects we have listed so far are resulting from the implementation of the CLES, which we see as a lever for effective change. We will now analyse the way change occurred.

## 3 Systemic Analysis of the CLES Organisation

Our approach is that of complexity, as presented by Morin (1988), and completed by the research led by Gélinas & Fortin (1996) who introduced the concept of “enovation”, according to which change arises from the actors involved in the certification. This concept matches the idea that the actors are at the origins of emerging novelty within a specific context. The strategy is that of “emerging change” and follows a bottom-up movement which implies that it is because of the appropriation by the actors (at a micro level), who have to deal with constraints (at meso and macro levels), that a change in the system is made possible. This concept is in opposition to the “innovation” concept that follows a top-down dynamic, which does not systematically allow positive change. The synergy of their theories will allow us to propose a model of positive washback for the implementation of a test in university context.

### 3.1 The Bailey Model

Kathleen Bailey counts among the researchers who took forward the research in the field of language testing, especially with the washback model she proposed and in which she develops the tripartite distinction made by Hughes (1993) between the participants, the processes and the products.

Bailey (1996, p. 264) breaks the linearity of the process by incorporating the notion of retroactivity. She also adds the entry of the impact a test can have on researchers. In the second part of the triptych she proposed, she indicates that the processes may concern changes in methodology among teachers or the use of strategies related to the passage of the test. Though Saville (2000, p. 3) underlines the fact that the processes which allow the change do not appear in the various models of washback until then proposed, because “they were not understood nor well represented in the model”.

Bailey, at the international conference of ALTE (Bologna, 2017), proposed a revised version of her model integrating interrogation points in the processes part. This last version is of particular interest to us since our aim is precisely to highlight the processes of implementation of a positive washback in university context with the CLES. To do this, and before we produce our model, we will proceed to the systemic analysis of the CLES by following the path of complexity.

### 3.2 Understanding the Process of Change with the CLES

Morin (1988) identifies eight different “avenues” of complexity to explain the behaviour of an organisation. Transferring his theory to the CLES organisation allows us to understand how change is made possible and effective.

The CLES is at the same time a centric (national coordination), polycentric (centres and hubs) and acentric (individuals) organisation. It is thus at the same time multiple and entire (5th avenue: unity and multiplicity). There is at the same time convergence towards a common project and divergence according to local contexts (6th avenue: contradiction between convergence and difference); some centres may choose to have their students pay for the certification and others not for example. Hence there can be unpredictable behaviours (1st avenue: unpredictability), which complicate the system (3rd avenue: complication) but which can bring out qualities some actors would not have been able to develop in the absence of the organisation constraints (7th avenue: constraints and emergence), such as the development of exam subjects for example. Hence, the actor is the product of the organisation but he is also contributing to its production (2nd avenue: transgression between the singular, the local and the universal). Thus, there is a strange relation established between order, disorder and the organisation (4th avenue: disorder and organisation). Last, implementing the CLES allows the teacher to reflect on and adapt his teaching posture. It also invites the student to play an active role in his learning, and become aware of his condition as a future citizen (8th avenue: reflexivity).

What we learn from this theory is that change was made possible following a bottom-up process, according to an “enovation” dynamics as introduced by Gélinas and Fortin (1996). The process of emergence is made possible by the creation of meaning and producing change as a support to the adaptation of practices, involving proactivity.

This systemic analysis allows us to try and complete Bailey’s model.

### 3.3 A Transferable Model for Positive Washback?

Our contribution to the model of Bailey is first that of the definition of the processes involved to allow positive washback. The identification of these processes allows us to identify new retroactive effects on the participants.

To allow positive washback the teacher has to be involved in the processes of examination, design and training; the student has to gain autonomy and evolve towards a new paradigm using language as a tool to perform communicative tasks. As an integrated tool, the management of the certification must be centralised to ensure that the produced effects and the assigned means are identical on the whole structure and for all the students. Last, by adopting a research-action posture, researchers can evaluate the quality of change and their results have a direct impact on the tool, in a retroactive movement.

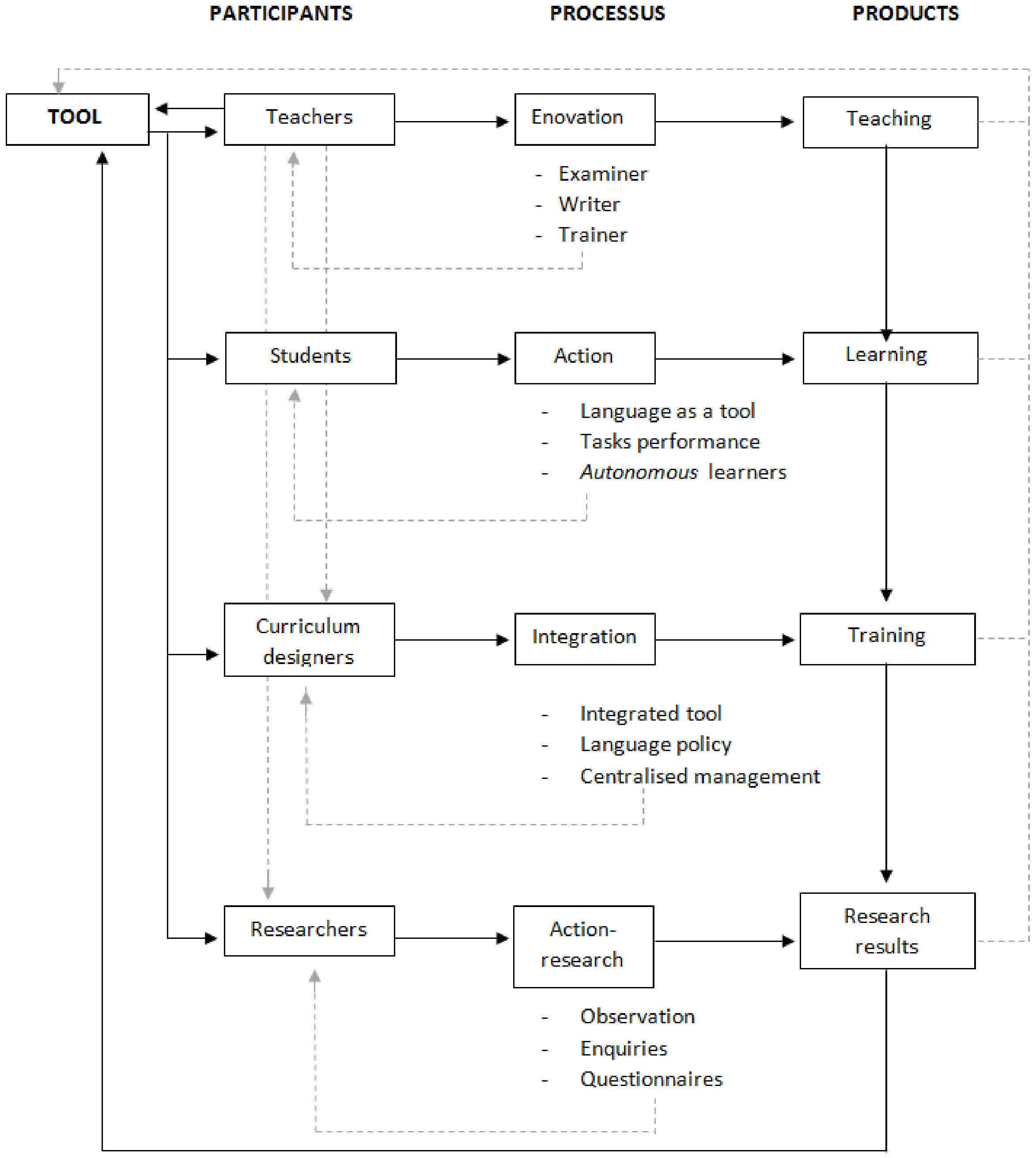


**Figure 1.** Proposed model of positive washback after Bailey's (1996)

## 4 Conclusion

Our analysis presents the CLES as a lever for positive washback within the context of higher education. Initially chosen as a possible answer to the LMD reform requirements, it proved to be an efficient tool in the service of a whole system, meeting micro, meso and macro level needs and issues. Moreover, at the institutional level, it appears that the CLES has proved to be a tool of equalisation, in the sense that it allowed a de-compartmentalisation of spaces, training and teachers. It also opened on an ascending de-stratification (it makes the actions of the different levels interdependent) of the institutional organisation, insofar as it allowed the professional and personal development of the teachers involved. Thus, we can assert that the CLES is a quality certification that generates quality.

## References

Alderson, J. C., & Wall, D. (1993). Does washback exist?. *Applied linguistics*, *14*(2), 115–129.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language testing, 13*(3), 257–279.

Fave-Bonnet, M. F. (1994). Quelle formation pédagogique pour les enseignants du supérieur. *Recherche et Formation, 15*, 11–34.

Gelinas, A., & Fortin, R. (1996). La gestion du perfectionnement des enseignants: formation-recherche auprès des directeurs d'établissements scolaires au Québec. In M. Bonami, & M. Garan. *Systèmes scolaires et pilotage de l'innovation. Emergence et implantation du changement* (pp. 115–145). Bruxelles: De Boeck.

Hughes, A. (1993). *Backwash and TOEFL 2000*. Unpublished manuscript, University of Reading.

Morin, E. (1988). Le défi de la complexité. *Revue Chimères, 5*(6), 1–18.

Poteaux, N. (2014). Les langues étrangères pour tous à l'université: regard sur une expérience (1991–2013). *Les dossiers des sciences de l'éducation, 32*, 17–32.

Saville, N. (2000). Investigating the impact of international language examinations. *Research Notes, 2*, 4–7.

# Instructors’ Perceptions of the Construct-Relevance of Language in the Assessment of Literature

**Sayyed Rahim Moosavinia**, Shahid Chamran University of Ahvaz, Iran
**Kioumars Razavipour**, Shahid Chamran University of Ahvaz, Iran

**Abstract:** One acute dilemma in testing literature in undergraduate English literature programs in EFL contexts is the extent to which language should constitute part of the target construct intended to be measured. In practice, language proficiency considerations seem to creep into assessment. The current study seeks to examine instructors' perceptions of the degree to which the quality of language forms part of the construct of testing literature. To this end, six literature professors from two departments of English and Persian were interviewed. Data analyzed via axial coding indicated that although language proficiency does feature in the assessment of literature in both the mother tongue and in testing literature in EFL, ideas are split as to the weight that must be given to language issues in such assessments. The study carries implications for high stakes external exams as well as achievement testing in undergraduate programs.

## 1 Introduction

It goes without saying that testing is (one of) the most challenging phase(s) in EFL contexts. When teachers reach this phase they usually cut it back. Nevertheless, it is difficult to imagine that the educational systems could afford to cut back on testing any further. Moreover, teachers and researchers are educationally and socially responsible to spread the consciousness about the significance of exams and the effects they have on teaching methods. Testing literature according to Paran (2010) has its own complexities and seems like being between Scylla and Charybdis (p. 143). As the philosophy behind teaching language and teaching literature is different, the policy behind the assessment of language and assessment of literature should also be different. Thus testing literature seems even a more daunting challenge.

The reason for teaching literature is more likely to give students insight into their personal lives and life in general. Students and even teachers are empowered intellectually and spiritually through the process of learning and teaching literature. Thus teachers should give tests that gauge similar understanding and insights. We should also control and make up for the state of testing literature which seems to lag behind in EFL testing. This is true about both high stakes exams and undergraduate/graduate achievement tests of literature. It is clear that assessment is a significant part of every teaching/learning activity and teaching literature is no exception.

Needless to say, tests affect and are affected by materials, curricula and teaching methodologies. Besides, it has been taken for granted that language and literary competence very often than not go together in both classrooms and tests. However, the two sides of this Janus-headed dilemma should fit into the whole picture of testing literature through language. Teachers’ experiences have showed that their perceptions of the construct-(ir)relevance of language to literary competence have been vaguely overlooked. However, it requires little attention to peep into classes and exam sessions, especially the corrected exam papers, to notice the admixture of literature and language.

Lack of devotion to literature testing has created a vast space to initiate literary tests and texts consistent with institutional guidelines rather than individual aspirations. Of course, “the objectives in literature teaching should form a harmonious whole" (Purves, 1986, p. 323). Therefore, educational systems should aim at producing students who are both knowledgeable and interested in literature, through the means of language. However, a recurrent question in the classroom is whether language proficiency affects assessment of literary competence and to what extent. Thus building a connection between teaching and testing literature in EFL contexts is highly significant as a degree of foreign language considerations seems to creep into assessment and affect instructors' judgment of students' literary competence.

## 2 Review of literature

It is often said that quantification and measurement would destroy beauty, which is an essential property of literature (Gaston, 1991). Whether this is the case or not, in the accountability era there is no escape from reporting outcomes in numerical narrative (ibid). Thus, assessing achievement in literature programs is inevitable. Yet, though teachers frequently assign numerical marks to students' work, assessing achievement in literary competence is easier said than done (Beach, 2014). In this regard, Paran (2010, p. 153) has identified six dilemmas for testing literature in EFL teaching: (1) whether testing is an external activity with a set of gate-keeping goals or an internal activity with a cluster of internal goals, such as individual growth and character development; (2) whether to teach language or teach literature; (3) to test literary knowledge or literary competence skills; (4) testing public literature knowledge (efferent reading) or personal appreciation of literature (aesthetic reading); (5) introducing genuine everyday oral tasks or formal non-specialist pedagogic tasks; and (6) teaching skills or teaching vocabulary about skills or metalanguage.

To make matters even more complicated, Purves (1986) opines that, like other types of language assessments (Fulcher, 2013), in assessing literature the pedagogical objectives should be borne in mind. Purves enumerates three purposes for teaching literature.

(1) Transfer of knowledge within literary/cultural texts of a group
(2) Training qualified readers and critics of such texts
(3) Promotion of personal empowerment by literary texts through the other two aims

Purves concludes that pedagogical objectives would bear on the test types and functions. Accordingly, when transfer of knowledge is the objective in teaching literature, tests need one to reflect and concentrate on questions which demand remembering and recalling. If the aim is training critics, tests of literature would be tests of skills. Finally, in situations where literature is taught to promote empowerment, assessing learners' and test takers' attitudes would be of primary importance (p. 323).

Whether circulation of literary knowledge in texts, preparing critical readers and critics of the texts or literary creativity and empowerment is in mind, at the end of the day, writers, readers and critics are all involved in language competence. Nevertheless, methods of teaching language are not sufficient for teaching literature. Language remains the medium for structuring human experience and imagination. Furthermore, literary devices and techniques foreground the literary language against the backdrop of the ordinary everyday language or even formal language. Thus Purves (1986) has clarified this question:

> Research indicates that the ratings of various aspects of performance are related to each other, but that raters aware of the relationships can make distinctions between the content and the form of a written or dramatic performance. For an overall grade in language arts, of course, teachers might want to combine the two, but for the literature aspect of the grade, the content is important. (p. 323)

On the whole, it really becomes a challenge to set goals for teaching literature. Indeed, very often various goals coexist during a literature or poetry course. The primary goal seems to be the immediate course work or course text(s). However, it is very common that teachers go beyond the syllabus and introduce, review or refer to other textbooks or materials. This issue is sometimes reflected in the type of assignments introduced by the teacher. Therefore, the idea here is to suggest that teaching literature is unique in that it easily connects to other texts and contexts. But what further complicates the question is to decide to measure the degree of the presence of language competence in testing literature. The main concern for the researchers here is to gauge the degree of consideration for language competence in testing literary content and competence.

Lastly, it is high time for teachers and faculty members to get engaged in literary courses and fill the vacuum projected by the nature of literary studies through “some sort of questionnaire or informal interview” (Purves, 1986, p. 323). Paul L. Gaston asserts, “Quantification and appreciation rarely coexist easily”, and “Measurement […] can be unkind to beauty” (Gaston, 1991, p. 1). Still to practically save the value of literary studies in a measuring and measured world, it is time to carry out this important task. However difficult it is for professionals in literary studies to perform such a task and resolve the dilemma, determination is always looming on the horizon to settle such problems.

## 3 Methods of the study

The study reported here is part of a larger study conducted in the Faculty of Letters and Humanities in Shahid Chamran University of Ahvaz, located in the southwest of Iran. A total of six literature instructors, all the three instructors from the department of English and three instructors selected from the Persian department comprised the participants of this study. The Persian literature instructors were selected among those whose cooperation with the study we could secure. All the interviewees were males and in the age range of 30 to 50.

To collect the required data, the six instructors participated in semi-structured interviews. The aim here was to tap on teachers’ ideas on the degree to which they see language proficiency as part of the construct of literary competence. To conduct the interviews, the first author met the participants at a time and place of their convenience. The interviews were semi-structured, that is, interviewees were required to adhere to the themes of the interview, which was on the extent they thought language proficiency must inform their literary competence evaluation decisions in the poetry courses they were teaching at the time. With their consent, their responses were audio-recorded via a mobile device, which were later transcribed for more in-depth analysis. In addition, notes were taken during the interviews to complement the insights gleaned from verbatim transcription. In total, 10 questions served as triggers for the elicitation of interview data from the participants. Data analysis was mainly informed by the questions raised during the interviews.

## 4 Results and discussion

Overall, Persian literature instructors appeared to believe that language competence and literary competence are overlapping constructs. They mostly emphasized language for literary competence; however, they deemed that it is far-fetched to rely on only written exams to test literary achievement. One of the participants maintained that “language and its related skills become the means for a better understanding of a literary text and acquiring literary insight.” Another instructor postulated that oral exams including the correct recitation of poetry could occupy a crucial place in the assessment of poetry courses. Language and literature always come together in teaching and testing. Language is the means for understanding literature and sometimes it becomes literature itself. To these instructors, the borders between language proficiency and literary competence are fuzzy, at best.

Less agreement or even consistency was observed among the English literature instructors. Two of the three participants placed special emphasis on language issues in assessing response to literature, perhaps mindful of the washback effect such a method might have for EFL learners. Language and literature, they admit, always come together in instruction as well as in testing. One instructor seemed to be on the same page with the Persian instructors, believing that language knowledge and literary competence are indeed the same. The third participant from the English department held a different attitude regarding the construct relevance of language to literary competence. He pointed out that marks should not be assigned to language issues in testing literature, poetry in this case, and that language proficiency issues should be treated in some other basic courses especially designed for promoting language skills.

In response to the question whether to allocate marks for fluency and literariness in writing about poetry in relevant exams, the six instructors reported assigned minimum weight to language problems such as punctuation and grammar. On the other hand, most instructors accepted that the students who enjoyed a better literary competence possessed higher level language proficiency. However, one of the English literature instructors held that the reverse is not true, indicating that higher language proficiency does not necessarily come with higher levels of literary competence.

## 5 Conclusion

Even with this small sample of participants, a few conclusions may be safely reached. First, to most participants, it is very challenging to demarcate response to literature between language and literary competence and that the two are the two sides of the same coin. Secondly, it dawned on us that testing literature in the mother tongue and testing it in a foreign language do diverge. More divergence was observed among the three English literature instructors. As noted earlier, one English instructor strongly adhered to the construct-irrelevance of language issues in testing literature. Finally, there was an obvious uneasiness among the participants of both languages to talk about the issue, seemingly lacking the pertinent vocabulary to elaborate on testing literature. The takeaway from this observation is that for the assessment of literature in both the L1 and the L2, instructors need to be made more conscious of the intricacies that are involved in the assessment of literature. A dose of assessment literacy is badly needed (Popham, 2006).

This study could barely scratch the surface of a complicated issue: assessing literature. Future research should delve more deeply into the components of literary competence across diverse literary genres. Another interesting line of inquiry is to see that training in assessment can engineer change in the way instructors go about assessing literary competence. Further, high stakes tests of literature deserve scholarly scrutiny, particularly for their consequential validity for literature education programs.

## References

Beach, R. (2014). Assessing responses to literature. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 85–101). John Wiley & Sons, Inc. DOI: 10.1002/9781118411360.wbcla017

Fulcher, G. (2013). *Practical language testing*. Routledge.

Gaston, P. L. (1991). “Measuring the Marigolds”: Literary Studies and the Opportunity of Outcomes Assessment. *The Journal of the Midwest Modern Language Association, 24*(2), 11–20.

Paran, A. (2010). Between Scylla and Charybdis: the dilemmas of testing language and literature. In A. Paran & L. Sercu (Eds.), Testing the Untestable in Language Education (pp.143–165). Bristol: Multilingual Matters.

Popham, J. W. (2006). Needed: A Dose of Assessment Literacy. *Educational Leadership, 63*(6), 84–85.

Purves, C. A. (1986). ERIC/CRC Report: Testing in literature. *Language Arts, 63*(3), pp. 320–323.

# Diagnostic Assessment: Incorporation of Assessment and Teaching in Foreign Language Education

**Dr. Hyunsoo Hur**, Defense Language Institute Foreign Language Center, USA

**Abstract:** This article introduces face-to-face diagnostic assessment practices adopted at the Defense Language Institute Foreign Language Center (DLIFLC), USA. Diagnostic assessment, within the framework of dynamic assessment, is used to gauge foreign language learners' proficiency levels, to identify learners' strengths and weaknesses, and to provide tailored instructions. Assessment exists in a continuum with instructions and DLIFLC adopts diagnostic assessment to assist learners reach higher levels of proficiency to better meet the government mission. The article discussed different stages involved in diagnostic assessment, how face-to-face diagnostic assessment is conducted institution-wide, and its significance in foreign language education.

## 1 Introduction

The choice over a certain type of assessment and the assessment procedure is often influenced by the objectives of the assessment, the assessment content and the assessment context (Shohamy & Inbar, 2006). Considering the needs, purpose, desired outcomes, and the extent of the population intended to be reached, different assessment formats take priority, be it traditional standardized testing or alternative performance-based. For example, the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) regards accuracy in the areas of syntax, morphology, lexicons, phonology, pragmatics and discourse as strong indicators of successful communication and offers both face-to-face and computerized tests. Test of English as a Foreign Language (TOEFL) also focuses on accuracy and provides indicators for foreign learners' competence in English. Alternative assessment also takes place in the instructional settings and examines learners' language performances and/or productions (e.g., portfolios, webpages) as indicators of accomplished learning. The similarity across these tests is that they identify what and how much the learners already know. Depending on purpose and use of the assessment, they may function as predictions of the future and/or focus on measuring the person's history as a learner instead.

At Defense Language Institute Foreign Language Center (DLIFLC) in Monterey, California, various assessment types, achievement, proachievement, proficiency (e.g., unit assessment, end-of-course test, OPI, Defense Language Proficiency Test, diagnostic assessment), are employed to measure language learners' level of proficiency. Among several types of assessment, diagnostic assessment is used institution-wide to better identify areas learners have not fully developed. This diagnostic assessment is adopted either online, computerized, or offline face-to-face. In this article, I will introduce offline face-to-face diagnostic assessment conducted at DLI. I will first examine the notion of dynamic assessment in relation to Vygotsky's sociocultural approach to mind (1978). I will then delve into diagnostic assessment, and how diagnostic assessment within the framework of dynamic assessment positions in the continuum with instructions, and functions to complement instructional practices.

## 2 Dynamic assessment

Static assessment or traditional types of standardized testing are interested in already matured abilities, and thus focus on a learner's past development. The examiner is expected to take a distanced and neutral stance and little or no feedback/mediation is provided during the assessment process.

Dynamic assessment has developed as an alternative to complement static assessment. It is used to gauge language abilities, intervene in learning, and record learners' growth (Anton, 2009). Conceptually based on Vygotsky's notion of sociocultural approach to mind, especially Zone of Proximal Development (ZPD), dynamic assessment is based on the view that our relationship to the world is mediated by physical and symbolic artifacts (e.g., books, computers, dictionaries, numbers, and language). Higher forms of thinking are socially and culturally derived, and through constant interactions with external resources, human cognitive development occurs intermentally and intramentally.

Vygotsky conceptualized two levels of developmental stage in his formulation of ZPD. One is actual developmental level where an individual is able to complete the task alone and potential developmental level where an individual can reach a level higher through assistance from external resources, such as an adult, a more advanced peer, or cultural artifacts. The intermental relationships with surroundings enable the child to access affordances (Van Lier, 2000) that are within social contexts. According to Lantolf (2000), the affordances are regarded as affordances only if they are within the individual's ZPD. Through collaborative interactions with affordances, cognitive development occurs, which very much depends on the type of mediation people have in their living activities.

Vygotsky (1978) argues that learning leads to development. An individual's ZPD should be discovered in order to fully understand the person's potential to develop (Poehner & Lantolf, 2003). In regard to language proficiency, Lantolf and Frawley (1988) argue that "proficiency is not a property of an individual but is a feature of the functional system formed between individuals and their unique, as well as shared, goals" (cited in Poehner & Lantolf, 2003, p. 4). Thus, when a performance takes place, it can be viewed as "a joint construction by the participating individuals" (Swain, 2001, p. 278). Thus to better identify a possible future further developed than the present, dynamic assessment in essence takes into account a potential level of development that could be reached by learners with external sociocultural assistance. Namely, depending on the learners' current level of development, individuals demonstrate different potential levels of development even with the same learning materials and instructions. Dynamic assessment regards assessment and instruction as "dialectically integrated as the means to move toward an always emergent (i.e., dynamic) future" (Poehner & Lantolf, 2003, p. 5).

Compared to traditional static types of testing, different dynamics occur in dynamic assessment. The examiner intervenes in the assessment process and positions themselves as a helper and supporter rather than playing the traditional role of dominant, neutral almighty. Assessment and instruction are united and the examiner intervenes as a teacher and diagnoses a learner’s cognitive and/or linguistic development. Often in the form of prompts, the examiner provides one-on-one assistance to the learner so that he/she can complete the assessment process which has been tailored to the learner’s individual developmental process, i.e., the learner’s ZPD. This approach to assessment is based on Vygotsky’s recognition that the diagnosis of the underlying sources of development is more important than documenting completed development (Davin, 2016). Vygotsky was mostly interested in “qualitative assessment of psychological processes and the dynamics of their development” (Anton, 2009, p. 579).

Dynamic assessment can be sub-categorized into interventionist and interactionist approaches (Poehner & Lantolf, 2003). An interventionist approach has more of a psychometrician orientation, thus is more geared towards quantifications of the data. An interactionist approach focuses more on qualitative assessment of psychological processes, thus tends to be more geared towards mediations through interactions. Diagnostic assessment conducted at DLIFLC has more of an interventionist approach. In the following section, diagnostic assessment is examined to see how its underlying ideas are adopted in real-life instructional contexts for diagnostic purposes.

## 3 Diagnostic assessment at DLIFLC

DLIFLC is a premier foreign language teaching institution within the US government. Designed to raise personnel working for the US government mission for national security, the institution comprises organizations intensely involved in various aspects of language teaching and learning, including language schools, divisions for assessment and testing, curriculum and material developments (traditional paper-based and technology-based), and teacher training. The institution also outsources various foreign language projects as well as developing internal projects to promote language proficiency. Twenty-three languages and two dialects are taught as resident courses in the Presidio of Monterey, California, and 98% of the faculty are native speakers of the foreign language of instruction.

Upon completion of respective language programs, students are required to take the Defense Language Proficiency Test (DLPT), which grounds on the US Interagency Language Roundtable (ILR). Since the graduates of DLIFLC are assigned to do real-life tasks related to national security, to better accommodate the government mission, the institution aims to have learners reach ILR level 2+ and 3, equivalent to the ACTFL proficiency scale “upper advanced” and “superior”. Leaver and Shektman (2002) stated that different types of language teaching methodologies should be applied to lower level learners and higher level learners. They also alluded to the fact higher level learners need more individualized tailored instructions as they have already developed idiosyncratic learning strategies and have unique needs and desires. As such, diagnostic assessment implemented institution-wide at DLIFLC seems to be a handy tool to identify areas that learners still need to develop for higher levels, as well as address each learner individually through one-on-one interactions between the evaluator/teacher and the learner, and connect assessment with instruction.

At DLIFLC, diagnostic assessment is regarded as an assessment tool that seeks to identify what a learner can do, what a learner cannot do, where the learner should be in their learning progress, and how to help the learner achieve their learning goals. It determines the gap between where the learner currently is and where he/she needs to be in order to target those skills required to achieve the target level. Its intent is to identify the strengths and weaknesses of the skills being assessed in order to make a clear assessment. While OPI and diagnostic assessment both establish a floor (what a learner can do) and ceiling (what a learner cannot do), OPI concentrates on completed learning and assigns the learner's proficiency level accordingly, and diagnostic assessment functions as formative assessment and orients towards providing a personalized individual development plan tailored to individual learner needs so that learners can use it as a tool for self-regulation.

Diagnostic assessment at DLI composes of three stages: 1. Pre-interview data collection, 2. Interview, and 3. Post-interview follow-up. Pre-interview data collection involves creating a learner and linguistic profile. Students complete a biographical questionnaire and share information on the family, previous DLPT scores, hobbies, traveling experience, education, previous foreign language learning experience, cultural awareness, current learning situation, etc. The learner and linguistic profile also includes personality type indicator, cognitive styles indicator, sensory preferences, motivation profile, and a writing sample in the target language. Personality type indicator uses a Myers & Briggs Type Indicator (MBTI) questionnaire and categorizes learner information through combinations of 8 basic personality types: Extraversion (E), Introversion (I), Sensing (S), Intuition (I), Thinking (T), Feeling (F), Judgment (J), and Perception (P). Cognitive styles indicator uses Ehrman & Leaver's (E & L) learning style questionnaire and categorizes cognitive styles into "synoptic" and "ectenic." Synoptic learners have a tendency to be global, impulsive, field sensitive, random and synthetic, whereas ectenic learners tend to be abstract field insensitive, reflective, deductive, sequential and analytic. Sensory preferences identify a learner's various sensory channels: visual, auditory, tactile, and kinesthetic. Motivation profile relies on a motivation survey. A writing sample in the target language requires the learner's sample writing on topics related to life history (e.g., daily life, travel experience) or future plans. The collection of writing samples is intended for a preliminary estimate of the learner's proficiency level through analysis of linguistic features (e.g., grammar, vocabulary, connectors). Based on the information learners provided in the questionnaires and the writing sample, the examiner/teacher creates a learner and linguistic profile.

The second stage interview is for the diagnosis of a learner's language proficiency. This is the stage where the underlying values of dynamic assessment are realized. The evaluator/teacher provides assistance to the learner but at the same time tries to gauge the level of proficiency. Using the ILR scale, three-skill interview, listening, reading and speaking, identifies the learner's current level of proficiency, the learner's weaknesses, and proximate level of development. The reading and listening interview relies on prepared texts in the target language with comprehension questions. The comprehension questions are designed to assess global tasks and functions, types of texts produced, lexical and structural control, delivery, and sociolinguistic competence. On average, about three to five texts are used but the evaluator/teacher can decide on the number of texts in accordance with the amount of data needed for collection. The speaking component follows a similar process to OPI, using probes to set threshold (a floor of sustained performance) and ceiling (limit where performance cannot be sustained) levels. Nevertheless, the concentration is more on identifying weaknesses that require treatment for future development.

The third stage consists of post-interview follow-up. Upon collection of all the data for the learner and linguistic profiles and the dynamic assessment interview, an Individual Learning Development Plan is created, followed by sharing of results with the instructional team and providing individualized feedback to the learner for future development. The teacher keeps the learner's diagnostic assessment profile and Individual Learning Development Plan, continuously revisits the profile throughout the duration of instruction, and works one-on-one with students to address individual needs and promote growth.

## 4 Conclusion

Testing is commonly used for gate-keeping purposes and used to make predictions about the future (e.g., Hanson, 1993; Shohamy, 2001). "Past-to-present models of development are typically employed" (Poehner & Lantolf, 2003) in such situations based on assumptions that the future and the present are equivalent. The purpose of dynamic assessment is to fill up the gap in the traditional notion of testing.

Dynamic assessment's goal is "to measure, intervene, and modify behaviors and to document the process of learning" (Anton, 2009, p. 579). Dynamic and diagnostic assessments have taken various forms and have been developed by different universities and organizations around the world, especially with online versions. The process and format of face-to-face diagnostic assessment adopted and developed at the DLIFLC could be another useful tool that could complement standardized testing as well as online versions of diagnostic assessment.

## References

Anton, M. (2009). Dynamic assessment of advanced second language learners. *Foreign Language Annals, 42*(3), 576–598.

Davin, K. (2016). Classroom dynamic assessment: A critical examination of constructs and practices. *Modern Language Journal, 100*(4), 813–829.

Hanson, F. A. (1993). *Testing testing: Social consequences of the examined life*. Berkeley, CA: University of California Press.

Lantolf, J. (2000). Introducing Sociocultural Theory. In J. Lantolf (Ed.), *Sociocultural Theory and Second Language Learning* (pp. 1–26). New York: Oxford University Press.

Lantolf, J. & Frawley, W. (1988). Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10*, 181–195.

Leaver, B. & Shektman, B. (2002). Principles and practices in teaching superior-level language skills: not just more of the same. In B. Leaver & B. Shektman (Eds.), *Developing Professional-Level Language Proficiency*. New York: Cambridge University Press.

Poehner, M. & Lantolf, J. (2003). *Dynamic assessment of L2 development: Bringing the past into the future*. CALPER Working Paper Series, No. 1. Pennsylvania: Pennsylvania State University, Center for Advanced Language Proficiency, Education and Research.

Shohamy, E. (2001). *The power of tests*. Mahwah: Lawrence Erlbaum.

Shohamy, E., & Inbar, O. (2006). *Assessment of advanced language proficiency: Why performance-based tasks?* CALPER Working Paper Series (CPDD 0605). Pennsylvania: The Pennsylvania State University, Center for Advanced Language Proficiency Education and Research.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*, 275–302.

Van Lier, L. (2000). From input to affordance: Social-interactive learning from an ecological perspective. In J. Lantolf (Ed.), *Sociocultural Theory and Second Language Learning* (pp. 245–259). New York: Oxford University Press.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge: Harvard University Press.

# Reconsidering The Impact of Language Assessment on Language Learning and Teaching: A Survey on an Italian Examination for Young Learners

**Paola Masillo**, University for Foreigners of Siena, Italy
**Carla Bagna**, University for Foreigners of Siena, Italy
**Sabrina Machetti**, University for Foreigners of Siena, Italy

**Abstract:** This paper aims to investigate the effect of the multilingual competencies and the maintenance of L1 of young learners of ISL in their performances. We conducted a survey to reflect on the multilingual competencies of young learners of ISL and on their resulting linguistic, communicative and educational needs. A questionnaire and two examination booklets were administered, respectively of CEFR levels A1 and A2, to a sample of young learners made of ISL learners and Italian native speakers. The final objective is to reflect on the test construct validity and its appropriateness for its purpose and context of use.

## 1 Introduction

The current paper aims at analysing the concept of impact as an extension of the notion of washback related to fairness and ethicality (Cheng, 2005; Green 2007; Tsagari 2011; Wall, 1997). In particular, the study concerns the effect and consequences a test can have beyond the classroom and immediate learning context.

The educational context mentioned in the current study is the Italian one. The researchers started from the premise that in the last decades, the presence of foreign children in the Italian schools has caused an educational emergency in which learning Italian and producing suitable materials for the teaching and assessment of communicative language competence in Italian represent the basic need (Bagna, Barni, & Machetti, 2004).

The main aim of the study is focusing on the development of assessment tools in Italy for foreign children and investigates, firstly, the contribution of language background to the related performance of both immigrant and non-immigrant students using an Italian language test for young learners of Italian as a second language. Secondly, the impact of an Italian language test for young learners on L1 and L2 students was analysed by exploring differences in participants' test outcomes compared to their educational achievement in Italian and Mathematics (Fox & Cheng, 2007).

The two main research questions consider, on one hand, whether these L1 and L2 test-takers' performances differ; and on the other hand, investigate whether the potential performance gap between immigrant and non-immigrant students can be closed. The answers to the two research questions proposed here will be useful to offer an early evaluation on the impact that such language assessment could have on the educational context, improving objectives and outcomes transparency, and increasing students' motivations and teacher accountability.

## 2 Overview of the study

In recent years, different national and international studies of education have often shown that the performance of immigrant students is substantially lower than that of non-immigrant students. As reported by Marks (2005), there is a range of explanations why immigrant students generally perform less well than other students. Those differences in performance are traditionally accounted for by socio-economic factors, sociocultural factors, and schooling factors.

As PISA test outcomes suggest, less wealthy children with an immigrant background face enormous challenges at school: they need to quickly adjust to different academic expectations, learn (in) a new language, and shape a social identity that incorporates both their background and their adopted country of residence (OECD, 2015a).

Looking at the Italian schooling context, in the last year the percentage of non-Italian citizens among students reaches 9.2 % overall. The non-Italian citizen students born in Italy are 51.7% over all the foreign students (Idos, 2015; MIUR, 2015a).

National programs for student assessment have consistently confirmed the PISA test outcomes (OECD, 2015b), showing a performance gap between students with an immigrant background and non-immigrant students. The administration of the Italian program for national student assessment in Italy, carried out during the last school year 2013/2014, shows that immigrant students achieve significantly lower results compared with those of their Italian citizen equivalents (INVALSI, 2014). In particular, the performance gap is wider between native and first generation immigrant students, as the second generation students' outcomes are closer to the Italian students' ones (MIUR, 2015b).

Looking at the field of language assessment, the Italian Ministry of Education has developed an official document stating the guidelines for the reception and integration of foreign students (MIUR, 2014). As reported in the ministerial document, it is a priority that the school favours, with specific strategies and personalized pathways, a possible adaptation of programs for individual pupils, which ensures non-Italian students an assessment, taking into account as far as possible, their previous educational history, the achieved results, the characteristics of schools attended, and skills acquired. The theoretical aim is to guarantee an evaluation that does not tend to lower the required goals, but adapts the tools to implement that assessment.

However, those strategies and theoretical assumptions do not find a systematic and consistent translation into practices actually implemented. Comparative international reports reveal the lack in Italy of a systematic regulation at national level, hence the variety of local practices (European Commission, 2013; Koehler, 2013).

## 3 Methodology

Considering the main objective of the study is to investigate the role of multilingual repertoire and the maintenance of the language(s) of origin in the performance of foreign pupils, an information questionnaire was administered in the first phase to all participants, focusing on standard demographic data and linguistic background. In a second phase, we focused on the educational outcomes achieved by the sample of students in Mathematics and Italian after 6 months of school, paying attention to the position of young immigrant students among others. In a third phase, teachers administered two booklets, of CEFR Levels A1 and A2 respectively, to the sample of young students, both immigrant and non-immigrant ones.

The administration was carried out at the end of the first half of the school year, when immigrant students (first and second generation) have attended a language course of Italian as a second language of an hour per week or an hour every fortnight. The tests were given to all pupils involved in the study, both Italian and foreign at the end of the first quadrant, for a total of 47 participants.

The measure of student status used in this study distinguishes the country of birth of both students and their parents. We define the following three categories:

- first-generation immigrants (foreign-born students whose parents are also foreign-born)
- second-generation immigrants (students who were born in the country of assessment but whose parents are foreign-born)
- non-immigrant students (both students and their parents were born in Italy)

The study was carried out in the primary school and involved students attending the fourth grade.

The language test adopted is a large-scale proficiency test of Italian as a foreign language. The test format is made up of five skills (Listening, Reading, Writing, Speaking and Use of Italian). The decision to use a language proficiency test for these purposes finds a first justification in Saville (2000), according to which the evidence provided in the international language examinations may have an impact on educational processes and on society in general. Further justification came in a series of international studies on the washback of a language competence test on the educational context (Alderson & Hamp-Lyons, 1996; Green, 2003; Shohamy & Donitsa-Schmidt, 1996; Tsagari, 2009).

## 4 Findings and discussion

In this section devoted to discussing the results, only a selection of the most significant data is presented. The data we will discuss do not pretend to be generalisable, but they intend to offer a first descriptive picture of the context being analysed.

### 4.1 Sample characteristics

The sample includes 47 students, 8 of whom are of foreign origins. Their age ranges between 9 (n = 7) and 10 (n = 1). They are 6 male and 2 female. Two students were born in Italy; among the others, there are 2 candidates of Peruvian nationality, 1 Egyptian, 1 Moroccan, 1 Moldavian, and 1 Ukrainian. The period of stay in Italy, when the place of birth is different from Italy, is quite variable. The range is from a minimum of 6 months to a maximum of 7 years.

The questionnaire checked the use of language(s) in the school context in the 2 contexts outside school: the family and the interactions with friends. The data describes a situation where multilingualism is a widespread phenomenon at home, where the use of a language other than Italian is very common, so most of the communicative interactions took place both in Italian and in a language other than Italian. In interactions with friends, the Italian language is definitely prevalent in 7 cases out of 8. The exception is a student (first generation) who said he uses both Italian and Arabic.

The last section of the questionnaire addressed the language self-assessment in both Italian and participants' mother tongue, when different from Italian, through a Likert-type scale. Considering the mean values, there is generally positive assessment for oral skills in both languages (L1 / L2); whereas the written skills values are lower.

### 4.2 Student achievement in Italian and Mathematics

The comparison between non-immigrant students and immigrant students starts with the analysis of their educational outcomes, making use of the marks obtained in the first semester of the school year 2015/2016.

Our sample of students (n = 47) scored almost equally well in Italian and Mathematics. If we consider the mean values, the achievement gap between students with and without an immigrant background was equal to less than 1 score point out of a range of 10.

### 4.3 Language proficiency test

The most significant results come out of the booklet of Level A2. The analyses refer only to test-takers who took all the parts of each test, in order to guarantee a comparability among data.

#### *4.3.1 Reading test*

The Reading Skill test section consists of three different test types: Multiple Choice, Finding Information, and Text Reorder.

Look at the mean scores, immigrant students on the Reading test were scored around 2 points lower than non-immigrant students on a 12-point scale. The main differences appear in parts 2 and 3. A second important finding to discuss is the number of non-immigrant students who got the maximum score (12 points): 14 out of 34.

The performance of immigrant students, who frequently use their L1 different from Italian outside school, was not substantially worse in Reading than that of non-immigrant students. Consequently, an implicational relationship cannot be found between the use of a home language other than that of education and the performance on reading tests. The maintenance of L1 and its use within the daily interaction contexts outside school do not negatively affect the outcome of the tests (Bagna, et al., 2004; Elley, 1992; Pauwels, 2004).

Another hypothesis to justify the low difference among the scores and the low number of successful results for non-immigrant students could be the intensive and extensive test preparation in their classes that some immigrant students received (Fox & Cheng, 2007).

### *4.3.2 Use of Italian test*

The test section dedicated to the Use of Italian has as its main purpose the assessment of linguistic and pragmatic competence. The test format is a cloze focusing on articles, verbs, and vocabulary.

The performance of immigrant students was no worse or different from that of non-immigrant students. In each exercise the immigrant students showed mean scores similar to that of non-immigrant students.

The cloze test allows test-takers to turn the attention on different traits of linguistic and communicative competence (lexical, semantic, grammar and spelling, but also textual traits), which includes evidence not only linguistic, but also metacognitive (Oller, 1979; Taylor, 1953).

The results obtained here gave us the evidence to reflect on the test construct itself and on its appropriateness for the context of use and target population and on the potential gap between what can be valued as literacy on the test and what can be valued in classroom literacy practice (Fox & Cheng, 2007).

### *4.3.3 Written production test*

The Written Production test is made up of two parts; the first part consists of a short, basic description of events, past activities and personal experiences. The second part required to write a very simple personal letter.

The low scores obtained even by the two second-generation students represent an interesting outcome (scores of 7 and 9, respectively, out of 12).

This evidence obtained could suggest to us the implications of a multicompetence (Brown 2013) or multilingual competence (Shohamy, 2006) for language assessment, particularly when the traditional benchmark is based on monolingual native speakers and it stipulates criteria for correctness and accuracy.

## 5. Final considerations

The research gave us the chance to investigate the implications of the multilingual competencies and the maintenance of L1 of young learners for language assessment (Brown, 2013; Shohamy, 2006;). Secondly, we reflected on the impact of language assessment on test construct validity and its appropriateness for its purpose and context of use (Messick 1996; Saville, 2012).

The data gathered and discussed so far gave us the first evidence to reflect on the role of a different language in young learners' language repertoire. Its use cannot be interpreted, as it happens sometimes mistakenly, as a form of refusal or closure with respect to the language of the school nor does it have a negative influence on training and integration in general (Little, 2010). Studies show that the development of linguistic competence in the L1 (if different from the language of education) can even favour learning and good school performance (McDermott, 2008; Robertson, 2006; Ryu, 2004; Thomas & Collier, 2002).

Secondly, we underline the need for a multilinguistic approach to language learning, as described in the CEFR (Council of Europe, 2001), still lacking in the Italian school system (Bagna et al., 2004; Brown, 2013; Shohamy, 2006). Studies show the importance of recognising the bilingual or multilingual linguistic repertoire of pupils for their school success (Conteh 2012; McDermott 2008; Parke, Drury, Kenner, Robertson, 2002).

The last closing remark refers to the need to reflect on the validity of the assessment criteria adopted, since they focus on a standard language, where grammar and vocabulary play a central role. Language assessors could perhaps consider language switches not as a failure to hit the target language but rather characteristic of multicompetent language use (Brown, 2013; Shohamy, 2006). This encouraged us to reflect on the way to assess bilingual or multilingual learners as it cannot be the same as we assess students with one language (Shohamy 2006).

## References

Alderson J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing 13*(3), 280–297.

Bagna C., Barni M., & Machetti S. (2004). La certificazione per bambini nelle fasi iniziali del processo di apprendimento dell'italiano L2. In I. Tempesta & M. Maggio (Eds.), *Lingue in contatto a scuola. Tra italiano, dialetto e italiano L2* (pp. 43–53), Milan Franco: Angeli.

Brown A. (2013). Multicompetence and second language assessment. *Language Assessment Quarterly* 10(2), 219–235.

Cheng, L. (2005). *Changing Language Teaching Through Language Testing: A washback study*. Cambridge : UCLES/Cambridge University Press.

Conteh, J. (2012). Families, pupils and teachers learning together in a multilingual British city, *Journal of Multilingual and Multicultural Development* 33(1), 101–116.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Elley W. B. (1992). *How in the world do students read?*. Grindeldruck GMBH: Hamburg.

European Commission. (2013). *Study on educational support for newly arrived migrant children*. Strasbourg: European Union.

Fox J., & Cheng L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education* 14(1), 9–26.

Green A. (2003). *Test impact and English for academic purposes: a comparative study in backwash between IELTS preparation and university pre-sessional courses* Unpublished PhD thesis, Centre for Research in Testing, Evaluation and Curriculum in ELT, University of Surrey, Roehampton.

Green A. (2007). Washback to learning outcomes: a comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education, 14*(1), 75–97.

Idos (Eds.). (2015). *Dossier Statistico Immigrazione 2015*. Rome: Imprinting Srl.

INVALSI. (2014). *Rilevazioni nazionali degli apprendimenti 2013/2014 – Prove INVALSI 2014*. Rome:INVALSI.

Koehler C. (2013). *Sirius, Policy implementation analysis by national and educational agents and other stakeholders*. Bamberg: European forum for migration studies, Institute at the University of Bamberg.

Little D. (2010). *The linguistic and educational integration of children and adolescents from migrant backgrounds*. Strasbourg: Council of Europe.

Marks G. N. (2005). Accounting for immigrant non-immigrant differences in reading and mathematics in twenty countries. *Ethnic and Racial Studies* 28(5), 925–946.

McDermott, P. (2008). Acquisition, loss or multilingualism? Educational planning for speakers of migrant community languages in Northern Ireland. *Current Issues in Language Planning*9(4), 483–500.

Messick, S. (1996). Validity of Performance Assessment. In Philips, G. (1996). Technical Issues in Large-Scale Performance Assessment. Washington, DC: National Center for Educational Statistics.

MIUR. (2014). *Linee guida per l'accoglienza e l'integrazione degli alunni stranieri*. Rome: MIUR.

MIUR. (2015a). *Gli alunni stranieri nel sistema scolastico italiano A.S. 2014/15*. Rome: MIUR.

MIUR. (2015b). *Alunni con cittadinanza non italiana. Tra difficoltà e successi. Rapporto nazionale A.S. 2013/2014, Quaderni ISMU 1*. Rome: MIUR.

OECD. (2015a). Can the performance gap between immigrant and non-immigrant students be closed?. *PISA in Focus, 53*(7): 1-4.

OECD. (2015b). Can schools help to integrate immigrants?. *PISA in Focus* 57(11), 1–4.

Oller J.W. (1979). *Language tests at school*. London: Longman.

Parke, T., Drury, R., Kenner, C., & Robertson, L.H. (2002). Revealing invisible worlds: Connecting the mainstream with Bilingual children's home and community Learning. *Journal of Early Childhood Literacy, 2*(2), 195–220.

Pauwels, A. (2004). Language maintenance. In A. Davies & C. Elder (Eds.), *The Handbook of Applied Linguistics* (pp. 719-737), Oxford: Blackwell.

Robertson, L. H. (2006). Learning to read "Properly" by moving between parallel literacy classes. *Language and Education, 20*(1), 44–61.

Ryu, J. (2004). The social adjustment of three, young, high-achieving Korean-English bilingual students in kindergarten. *Early Childhood Education Journal 32*(3), 165–171.

Saville, N. (2000). Investigating the impact of international language examinations. *Research Notes,* 2, 4–7.

Saville, N. (2012). Applying a model for investigating the impact of language assessment within educational

contexts: The Cambridge ESOL approach. *Research Notes, 50*, 4–8.

Shohamy E. & Donitsa-Schmidt, S. (1996) Test impact revisited: washback effect over time. *Language Testing, 13*(3), 298–317.

Shohamy, E. (2006). *Language policy : hidden agendas and new approaches*. London: Routledge.

Taylor, W.L. (1953). Cloze procedure: A new tool for measuring reliability. *Journalism Quarterly*, 30. 415–433.

Thomas W.P., & Collier V.P. (2002). *A National Study of School Effectiveness for Language Minority Students' Long-Term Academic Achievement*. Berkely: Center for Research on Education, Diversity and Excellence, Berkeley.

Tsagari, D. (2009).*The Complexity of Test Washback: An Empirical Study*. Frankfurt am Main: Peter Lang GmbH.

Tsagari D. (2011). Washback of a high-stakes English exam on teachers' perceptions and practices. In E. Kitis, N. Lavidas, N. Topintzi, & T. Tsangalidis (Eds.), *Selected papers from the 19th International Symposium on Theoretical and Applied Linguistics (ISTAL 19)* (pp. 431–445), Thessaloniki: Department of Theoretical and Applied Linguistics, Aristotle University of Thessaloniki, School of English, Aristotle University of Thessaloniki.

Wall D. (1997) Impact and washback in language testing. In C. Capham & D. Corson (Eds.), *Encyclopedia of language and education volume 7: Language testing and assessment* (pp. 291–302), Dordrecht: Kluwer Academic Publishers.

# Investigating Scoring Procedures in Language Testing

**George S. Ypsilandis**, Aristotle University of Thessaloniki, Greece
**Anna Mouti**, University of Thessaly and Aristotle University of Thessaloniki, Greece

**Abstract:** One among the main concerns of language testers in the design and implementation of tests is to select the method of scoring for the tool used to perform the evaluation. This attribute indirectly reveals the tester's ethical beliefs and personal stance in testing pedagogy. This study challenges the typical 1-0 method of scoring in Multiple Choice Tests (MCT) and implements a polychotomous partial-credit scoring system in official tests administered for the Greek State Certificate of Language Proficiency (GSCLP). The MCT items chosen were completed by a total of 1,922 subjects in different levels of the GSCLP test. Results clearly indicate that this scoring procedure provides refined insights to students' interlanguage level and enhances sensitivity in scoring procedures without jeopardising test reliability.

## 1 Introduction

Fairness is considered a "fundamental concern' in language testing, although 'describing this has proven elusive" (Bachman & Palmer, 2010, p. 127). Among the characteristics of fairness in language testing that have been discussed are those reported by Kunnan (2004): a) absence of bias, b) equity of access, c) validity of test scores, d) administration and e) impact. Bachman & Palmer (2010, p. 128) also suggested: f) equitable treatment of test-takers in the testing process, g) equality of testing outcomes for different groups, and h) equity in opportunities to learn the content that is measured in an achievement test. These tactics, mostly of a political aspect about testing, show that fairness covers a very large field of conceptualisation and the researchers/testers may deal with different characteristics, relevant qualities and perhaps measureable attributes. Some of those are related to test impact (e.g. g), planning (e.g. a+b), administration (e.g. d), or to the method of scoring (e.g. a), while one with general policies in education (e.g. h). It becomes apparent that fairness is a complicated issue and cannot be attributed to a test by a yes/no answer and thus "the best way to ensure test fairness is to build fairness into the development, administration, and scoring processes" (Zieky, 2002, p. 2). Despite its complexity, fairness is a fundamental characteristic of language testing because "irrespective of whether language assessments are used appropriately or inappropriately, they serve as both door-openers and gatekeepers" (Bachman and Purpura, 2008, p. 456).

One of the item formats commonly used in language tests, by which decisions are made, is the Multiple Choice (MC) which typically requires a selected response from among the choices provided. Two types are typically identified in MC selected response tests (Bachman and Palmer, 1996, pp. 202-203); best answer for the task types "in which the test-taker is expected to choose the best answer from among the choices given" (p. 202), and correct answers "which implies that there is only one correct answer in the world, and that this answer is among the choices provided". In the selected response best answer type, the level of item difficulty is defined by the quality of the distractors and the plausibility of the synonymous options. Arguably, a MC item could be constructed either with the correct answer standing out of the other options (which are equally and totally wrong) or with a less transparent correct answer and more plausible alternative options to distract the test-takers. In MC tests of single correct answer type, the scoring pattern 1-0 or 1-(-1) is usually followed. In MC tests of best answer type however, the above pattern may be found rather insensitive and probably unfair, as those test-takers who select a closer synonym against those who select a totally irrelevant answer are not rewarded.

This study lies within the area of test scoring procedures and applies to all test-takers alike. An experimental polychotomous partial credit scoring system is implemented and compared with the traditional dichotomous scoring procedure. It is hypothesised that this scoring system may provide a more refined score of the test-takers' performance and thus mirror his/her interlanguage stage. This sensitive scoring approach is expected to increase test reliability without jeopardising test results. By that respect fairness may be served.

## 2 Scoring and interpreting the test scores

Test results are calculated to produce some form of final score for each test-taker. A common method of scoring, provided there is no item weighting, is to assign one point to correct responses and zero points to the wrong ones. In particular, in selected response MC items this is considered the norm, while in constructed response items (e.g. gap-filling) partial credit scoring is also considered as an option. Lau, Lau, Hong & Usop (2011, p. 101) state that "the recognition of partial knowledge leads to the belief that a student's level of knowledge falls on a continuum ranging from full knowledge to full misconception." The authors review various scoring methods to credit partial knowledge: confidence weighting, elimination testing, subset selection testing, probability measurement, answer-until-correct, option weighting, item weighting, rank ordering the option, and partial ordering. Typically, distractors are equally weighted with 0, in 0-1 scoring, although there is differential information in them (Haladyna's, 2004, p.219 "differential distractor functioning"). Consequently, differentially attractive distractors could provide the basis for improving scoring of item responses as they could be differentially weighted according to their approximate correctness (see also Method below).

## 3 Method

The study's research design follows Tsopanoglou, Ypsilandis & Mouti's (2014), and Mouti, Ypsilandis & Tsopanoglou's (2013) studies where "option weighting" was used by awarding scalable points for choosing each MC option/answer/distractor. The option weighting approach may be implemented where MCQs contain distractors that are somewhat correct but not the best choice. This "weighting approach" is examined empirically by rewarding with partial credit scoring the test-takers who avoid selecting the totally irrelevant options in (polychotomous) MC items and choose a wrong although plausible option.

### 3.1 Participants

Two types of participants were engaged. The first group consisted of 4 native speakers/teachers and 2 proficient and experienced teachers of Italian (judges, from here on). Results from these judgements are presented below. The second group involved 1,922 test-takers who completed three Italian language tests in official settings (400 test-takers at A1–A2, 1,294 at the B1-B2 and 228 at the C1 levels). The L1 of the test-takers was Greek.

### 3.2 Materials

Data were collected from the Greek State Language Examinations for the Italian language (official tests in official authentic settings). The entire official test for each level included 4 papers (one for each macro skill): Speaking, Writing, Listening and Reading and Language Awareness. The study examined tests from Reading Comprehension and Language Awareness papers, from where a total of 53 dichotomously scored MC items (study sample) were extracted with 3 possible answers (1 correct and 2 wrong): 10 at the A1–A2, 15 at the B1–B2 and 27 at the C1 levels. The SPSS statistical package was used for test analysis.

### 3.3 Design, procedure and scoring system

In the study sample, polychotomous patterns and option weights were determined by the judges who ranked choices in a Likert scale, i.e. correct, very plausible/plausible and totally irrelevant/wrong. The polychotomous items were corrected with two modes of scoring: a) a traditional Dichotomous Scoring Method (DSM) where one (1) point is assigned for the selection of the correct answer and zero (0) points for all other choices, and b) a polychotomous scoring proposal (herewith Experimental Scoring Method, ESM) where one (1) point is provided for the correct answer, half a point (0.5) for the selection of the very plausible/plausible alternative and zero (0) points for the selection of the totally irrelevant/wrong answer.

## 4 Data analysis-scoring procedures

In the 53 items that were examined, divided in 5 testlets (sets of items), 67% (36 items/ Facility Index = 0.65) followed a dichotomous pattern and 32% (17 items/Facility Index = 0.37) a polychotomous one. It should be pointed out at this stage that judgments were not unanimous in all cases. In 11 items (25%), the polychotomous pattern was confirmed by the judges while in 6 items the expert judges were not able to trace the correct answer (being distracted themselves) and therefore these items were also examined and included in this category. These judgments were examined and verified empirically in relation to the item analysis results and the distractor analysis: in 11 out of 17 polychotomous items, the very plausible answer/option was the one with the highest choice mean/percentage compared to the correct answer. In addition, the average choice means, for both the correct and the plausible answers, were almost the same, although it would have been expected for the correct answer choice mean to be higher (correct answer choice mean: 0.42, plausible answer choice mean: 0.44).

The items found to have a polychotomous pattern were scored with both the DSM and ESM, while the ones where no polychotomous pattern was identified by the experts were only scored in the traditional way. Results from the different scoring procedures were compared and statistical analyses with SPSS followed to offer insights in terms of correlations and differences between the scoring procedures.

A Level: 1st set of items: 10 MC ITEMS

(Mean = 5.92, SD = 2.14, Alpha = 0.56)

Option weighting and item analysis verification was attempted following the expert's judgments. All the experts declared that all the above items were single correct answer, although the correct answer for the first item was not selected by all the experts. The specific set of items was not further examined with a polychotomous scoring method (there was only one item traced that presented traces of diferentiality). It may be argued that in lower levels, degrees of incorrectness and polychotomous patterns cannot be easily applied as this would increase significantly test difficulty.

B Level: 1st set of items: 7 MC ITEMS

(Mean = 4.12, SD = 1.60, Alpha = 0.46)

All judges found that all the above items were single-correct answer items. The specific set of items was not further examined and the polychotomous scoring was not implemented, similarly to the above.

B Level: 2nd set of items: 9 MC ITEMS

(Mean = 3.39, SD = 1.54, Alpha = 0.22)

Experts recognised a polychtomous pattern in 5 items, which included a semi-correct/plausible answer. These polychotmous items proved to be more difficult to answer than the dichotomous items as indicated by the Facility Index (Dichotomous FI = 0.47> Polychotmous FI = 0.32).The statistical analysis revealed that in 4 items the plausible distractor was chosen by more test-takers instead of the correct answer. In 3 of those items the selection coincided with the one provided by the expert judges as correct! (The selection distracted the judges as well).

The scores were altered when the Experimental Polychotomous Scoring (EPS) was implemented. In particular, the Facility Indexes were increased and the differences were statistically significant: Mean TDS = 3.38, Mean EPS = 4.5. In order to investigate reliability of the ESM, the Pearson r correlation coefficient was employed (examines the relationship among variables) to compare the independent variables in twos. Bachman (2004) proposes this test to investigate relationships among different sets of test scores. This revealed that the two scoring procedures do indeed exist in a strong linear relationship to each other. In more detail the value between TDS and EPS is r = .937 (p ≤ .001) and correlation is significant at the 0.01 level (2-tailed). Thus, test results are not jeopardised. Furthermore, a paired-sample T-test showed significant differences between the two scores t = 72.941, df = 1.293 (p ≤ .001) which supports the alternative hypothesis. Observable differences are not explained by random variation and thus the EPS offers a more sensitive scoring statistically different from the TDS.

C1 Level: 1st set of items: 12 MC ITEMS

(Mean = 7.36, SD = 1.84, Alpha = 0.41)

A polychtomous pattern at 6 items was identified by the expert judges. These also proved to be more difficult to answer than the dichotomous items as indicated by the Facility Index (Dichotomous FI = 0.67> Polychotmous FI = 0.39). In 3 items the distractor was chosen by most test-takers instead of the expected/correct answer. In 2 of these 3 cases the subjects' erroneous selection again coincided with the one selected by the expert's judgments! Apparently, the distractors were good, enough to mislead the native judges as well. Implementing the EPS, the scores were altered, the Facility Indexes were increased and the differences were statistically significant: Mean TDS = 7.36, Mean EPS = 8.20. This confirms again that the two scoring procedures do indeed exist in a strong linear relationship to each other. Pearson value between TDS and EPS is r = .0,966 (p ≤ .001), and correlation was found significant at the 0.01 level (2-tailed). A paired-sample T-test again showed significant differences: t = 26,215, df = 228 (p ≤ .001).

C1 Level: 2nd set of items: 15 MC ITEMS

(Mean = 8.41, SD = 2.41, Alpha = 0.49)

Experts recognised a polychtomous pattern at 5 items, which again proved to be more difficult than the Dichotomous items as indicated by the Facility Index (Dichotomous FI = 0.73> Polychotmous FI = 0.40). The statistical analysis revealed that in all 5 items, the plausible distractor was chosen by most test-takers than the correct answer and once again their selection coincided with the expert's judgments. Implementing the EPS system again altered the scores, the Facility Indexes were increased and the differences were statistically significant: Mean TDS = 8.22, Mean EPS = 9.46. Once again the two scoring procedures proved to exist in a strong linear relationship to each other, as the Pearson value between TDS and EPS was r = .967 (p ≤ .001) and correlation was significant at the 0.01 level (2-tailed). Similarly to the B level results above, a paired-sample T-test showed significant differences: T-test: t=29,509, df=228 (p ≤ .001) which again supported the alternative hypothesis.

## 5 Conclusion

Our hypothesis has been adequately supported by the evidence. In particular: a) the partial credit polychotomous scoring implemented has provided the expected refined understanding of the test-taker's language knowledge and b) test reliability was not affected, as the two scoring procedures were found to be in a strong linear relationship to each other in all cases. These findings support the results of our preliminary study (Tsopanoglou et al., 2014).

In norm-referenced situations the increase at the level of scores may not have had significant impact as the test-takers' ranking remained the same (high Pearson correlations). However, in criterion-referenced situations "where there exists a predetermined criterion for the students to meet, low scores would hurt those at the borderline" (Farhady, 1996, p. 222). It is here that our EPS would have a significant impact (supported by statistically significant T-tests).

Dichotomous items were easier than the polychotomous to answer, as the correct option in the former becomes transparent. Bachman and Palmer (1996, p. 202) indicated: "an item would be significantly more difficult if the options were closer in meaning because that would make identifying the correct answer more demanding for the test-taker". Polychotomous patterns have been traced at the higher B2–C1 and not at the lower A1–B1 levels (the higher the more) as these often are analogous to level difficulty.

Finally, it is our belief that the EPS method, adopted in our study, may provide a more complete view of the interlanguage stage of an individual and thus it contributes to fairness and score accuracy, particularly for those test-takers who show high level of target language awareness (by choosing a plausible answer and not a totally irrelevant option, through inferencing). In support of this claim, Bachman & Palmer (1996, p. 205) recommended that test-takers should be encouraged to make informed guesses and that "this should be rewarded, preferably through partial credit scoring".

## Further Reading

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2*, 79–96.

## References

Bachman, L. F. (2004). *Statistical Analysis for Language Assessment*. Cambridge: Cambridge University Press.

Bachman, L. F. & Palmer, S. A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bachman, L. F. & Palmer, S. A. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Bachman, L. F., & Purpura, J. E. (2008). Language assessments: Gate-keepers or gate- openers? In B. Spolsky & F. M. Hult (Eds.), *The Handbook of Educational Linguistics* (pp. 456–468). Oxford: Blackwell Publishing.

Farhady H. (1996). Varieties of cloze procedure in EFL Education. *Roshd Foreign Language Teaching Journal*, *12*, 28–37.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah: Lawrence Erlbaum.

Kunnan, A. (2004). 'Test fairness' In M. Milanovic & C. J. Weir (Eds.), *European Language Testing in a Global Context* (pp. 27–48). Cambridge: UCLES/Cambridge University Press.

Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge,and misconceptions in multiple-choice tests. *Educational Technology & Society*, *14(4)*, 99–110.

Mouti A., Ypsilandis G., & Tsopanoglou, A. (2013). *Investigating fairness in multiple-choice tests*. Paper presented at the 10th International Conference in Greek Linguistics, Komotini.

Tsopanoglou A., Ypsilandis G., & Mouti A. (2014). Piloting a polychotomous partial-credit scoring procedure in MC tests. *Language Learning in Higher Education, Journal of the European Confederation of Language Centres in Higher Education*, *4*(1). Retrieved from http://www.degruyter.com/view/j/cercles.2014.4.issue-1/cercles-2014-0004/cercles-2014-0004.xml

Zieky M. (2002). Ensuring the fairness of licensing tests, Educational Testing Service, *CLEAR Exam Review*, *7*, (1), 20–26.

# A Comparative Study on the Washback of CET-6, IELTS and TOEFL iBT Writing Tests: Evidence from Chinese Test-takers' Perspectives

**Xiangdong Gu**, Chongqing University, PR China
**Yue Hong**, Chongqing University, PR China
**Chengyuan Yu**, Chongqing University, PR China
**Tarun Sarkar**, Chongqing University, PR China

**Abstract:** Washback refers to the influence of test on teaching and learning. CET-6 (The College English Test Band 6 in China), IELTS and TOEFL iBT tests are all large-scale, high-stakes examinations of English as a foreign/second language. Based on the models of washback (Hughes, 1993; Xie, 2010) in language assessment and expectancy-value theory (Jacob & Eccles, 2000) in psychology, the present study compares the washback of CET-6, *IELTS* and TOEFL iBT writing tests on Chinese test-takers from two perspectives: washback on Chinese test-takers' perceptions and washback on their test preparation processes. A quantitative approach (questionnaire surveys) was adopted primarily in this study. Findings indicate that there is more similarities between IELTS writing test washback and TOEFL iBT writing test washback on Chinese test takers than CET-6 writing test. The washback of IELTS and TOEFL iBT writing tests includes more intense preparation activities that take place over a much longer period of time than for the CET-6 writing test. It is hoped that this study merits further investigation into writing test washback on learners in the Chinese context and beyond.

## 1 Introduction

This paper is a comparative study on the washback of CET-6, *IELTS* and TOEFL iBT Writing tests by collecting evidence from Chinese test-takers' perspectives. China has seen growing numbers of students taking these three tests in recent years. CET-6, *IELTS* and TOEFL iBT are large-scale, high-stakes and standardised English language tests, whose washback of the three tests is noteworthy to researchers. Writing being an important part in language learning and assessment is an integral part of all the three tests, and Chinese test-takers' poor writing performance in the three tests necessitates us to investigate and compare the washback of the writing sections.

The CET-6 writing test consists of one task of 30 minutes, accounting for 15% of the total score. *IELTS* writing test has two tasks of 60 minutes, accounting for 25% of the total score. TOEFL iBT writing test has two tasks of 50 minutes and the total score of writing is 30 marks, accounting for 25% of the whole score.

So far, many researchers have given their own definitions of washback, but almost all definitions contain the core concept of the influence of tests on teaching and learning, which is also adopted in the present study. In the literature review, we found previous empirical studies seldom specifically targeted at the washback of the writing test, and rarely focusing on the comparison of the washback of different tests and attached less importance to the washback on students than on teachers. Therefore, this study comparatively explores the washback of the three writing tests (CET-6, *IELTS* and TOEFL iBT writing) on test-takers.

The theoretical framework for the present study incorporates Hughes's Participants-Processes-Products washback model, Xie's washback model on learning, and Expectancy-value theory (Jacob & Eccles, 2000). The study intends to address two research questions:

Q1: What are the similarities and differences concerning test-takers' perceptions toward CET-6, *IELTS* and TOEFL iBT writing tests?

Q2: What are the similarities and differences concerning test-takers' test preparation toward CET-6, *IELTS* and TOEFL iBT writing tests?

## 2 Methodology

As for the methodology, this study mainly adopts a quantitative approach and employs questionnaire surveys to collect data. Three questionnaires (the CET-6 writing questionnaire (WQ), *IELTS* WQ and TOEFL iBT WQ) were prepared based on scales in the previous studies, findings of empirical washback studies, test-related official documents (e.g., test syllabus, official guidebook, preparation planner), expert consultation, test-taker interview data, and pilot study data. The finalised three questionnaires share the same constructs and have three sections. The first section focuses on the background information of the subjects, like gender and university. The second section aims to study the washback on the test-takers' perceptions, including five types of perceptions:

(1) Perception of test uses, including achievement and instrumental test uses.
(2) Perception of test design, including perceived assessed writing abilities and evaluation of test design.
(3) Self-concept of one's ability, including perceived non-writing and writing abilities.
(4) Task expectancy, including self-efficacy (test-takers' confidence on the test) and test results expectation.
(5) Subjective task values, including perceived positive and negative washback.

The third section concerns how test takers prepare for the tests. Five types of test preparation practices, we define are test preparation management, drilling (mass practice), memorisation, language skill development and social affective strategies. The other aspect we examine is test-takers' time investment on the preparation.

The formal questionnaire data collection started in June 2016 and ended in August 2016. Both paper-based and internet-based Chinese questionnaires were used to collect the data.

To minimise the regional differences of test-takers in the three questionnaire surveys, only those who answered the questionnaires with Chongqing Internet Protocol were retained in the current study. In total, we kept 106 CET-6 cases, 85 I*ELTS* cases and 73 TOEFL iBT cases.

Descriptive analysis on the demographic information of the subjects was conducted first. The subjects in the three questionnaire surveys share a similar age and have a relatively balanced coverage and proportion of gender, university, major, educational background and test-taking experience, indicating a high homogeneity of their background. Moreover, exploratory factor analyses and reliability analyses show that the scales in the three questionnaires have high construct validity and reliability. As per the results of exploratory factor analyses and reliability analyses of the perception scales in the questionnaires, the extracted factors are consistent with the intended constructs in each scale and the reliability of each scale is high. After factor analyses on the perception scales, 52 items in CET-6 WQ, 53 items in *IELTS* WQ and 53 items in TOEFL iBT WQ were retained. The results of exploratory factor analyses and reliability analyses of the preparation scales in each questionnaire reveals that the extracted factors are consistent with the intended constructs in each scale and the reliability of each scale is high. After factor analyses on the preparation scale, 18 items in CET-6 WQ, 17 items in *IELTS* WQ and 15 items in TOEFL iBT WQ were retained.

## 3 Results and discussion

### Q1: What are the similarities and differences concerning test-takers' perceptions toward CET-6, IELTS and TOEFL iBT writing tests?

For the first research question, we'll present the results and discussion from the five aspects: perception of test uses, perception of test design, self-concept of one's abilities, task expectancy and subjective task values.

Comparing the three tests in terms of perception of test uses, ANOVA analyses show that test-takers have both high perceptions of achievement test uses and instrumental test uses with no significant differences. Test-takers' high perceptions of achievement test uses in the three tests are consistent with the test developers' intention to promote English language achievement of the students, and test-takers perceive the valuable achievement test uses regardless of which test of the three they take. Moreover, test-takers' high perceptions of instrumental test uses indicate their utilitarian values of learning another language. The Chinese culture of learning is examination-oriented and has a utilitarian function, which can give rise to test-takers' strong perceptions of instrumental uses of the tests. In terms of perception of test design, the results show that, on the whole, test-takers think they need significantly stronger writing abilities in *IELTS* and TOEFL iBT writing than in CET-6 writing. Moreover, in all three categorisations of assessed writing abilities, namely content control, discourse organisation and language use, significant difference is found between CET-6 and *IELTS* and CET-6 and TOEFL iBT. The results further show that test-takers need significantly stronger writing abilities related to content control, discourse organisation and language use in handling *IELTS* and TOEFL iBT writing than CET-6 writing.

To compare the evaluations of the test design on the whole, test-takers give positive evaluations to *IELTS* writing the most and to TOEFL iBT writing the least in the three tests. Significant differences are found in the pair of the CET-6 and TOEFL iBT and the pair of the *IELTS* and TOEFL iBT.

Results of the comparison of three common items concerning the evaluation show that

(1) Test-takers think that scoring rubrics of *IELTS* and TOEFL iBT writing are significantly clearer than that of CET-6 writing. This finding is consistent with the previous findings

that CET lacks detailed scoring criteria on structure, content, examples, grammar and vocabulary.

(2) Though test-takers show their preference for paper-based tests over computer-based tests as a whole, TOEFL iBT test-takers show significantly less fondness on paper-based tests than CET-6 and *IELTS* test-takers. TOEFL iBT is an internet-based test and TOEFL iBT test-takers are more familiar with this test format, which can possibly decrease their favouritism to traditional paper and pencil test.

(3) Test-takers have a similar belief that CET-6, *IELTS* and TOEFL iBT writing scores can reflect their writing competence, based on which writing sections of the three tests have high construct validity to some extent. But more comparative studies on the construct validity of CET-6, IELTS and TOEFL iBT writing tests need to be carried out in the future to test this hypothesis.

In terms of self-concept of one's ability under the washback of the test, the results show that *IELTS* and TOEFL iBT test-takers not only perceive their overall language abilities higher than their CET-6 counterparts, but also their writing abilities and non-writing abilities more competent than CET-6 test-takers. One possible reason is that the greater challenge of *IELTS* and TOEFL iBT writing might spur *IELTS* and TOEFL test-takers to work harder on their preparation, which can improve their language abilities. Another reason for this result might lie in the samples. We achieve high homogeneity of the test-takers' demographic background, but their language abilities were not strictly controlled due to the difficulty of data collection.

In terms of test-takers' task expectancy, ANOVA analysis shows that *IELTS* and TOEFL iBT test-takers have significantly higher self-efficacy than CET-6 test-takers. Different perceptions of test design, specifically their perceived assessed writing abilities may contribute to the results. As discussed before, *IELTS* and TOEFL iBT test-takers have a better understanding about what is assessed, which can make them more confident.

Moreover, test-takers also show similarities and differences on their expectation on the test outcomes. Generally, a very low proportion of test-takers have low expectations on their results, and the majority of the test-takers have intermediate or high expectations. Most of the *IELTS* and TOEFL iBT test-takers have high expectations while CET-6 test-takers have intermediate expectations. Test-takers' English competence and their self-efficacy can contribute to the difference to some extent.

In terms of subjective task values, test-takers perceive that the three writing tests have strong positive washback on them; after comparing the common items, we found that the common top positive influences perceived by test-takers are all concerned with the positive impact on their overall writing competency rather than on other knowledge and skills. However, in general, they feel *IELTS* and TOEFL iBT writing tests have significant strong positive washback on them.

In addition, for negative washback, the present study finds that CET-6, *IELTS* and TOEFL iBT test-takers perceive that the tests have a similar intensity of negative washback on them with no significant difference. Additionally, all the highly ranked negative washback items seem to be related to test-takers' emotions, like making them feel anxious and dampening their enthusiasm to writing, which warrant that special attention should be paid to the negative washback on test-takers' psychological health.

**Q2: What are the similarities and differences concerning test-takers' test preparation practices toward CET-6, *IELTS* and TOEFL iBT writing tests?**

For the second research question, *IELTS* and TOEFL iBT test-takers have significantly more intense preparation practices than CET-6 test-takers, suggesting that in general, *IELTS* and TOEFL iBT writing tests impose more intense washback on test-takers.

However, for test-takers, they don't engage in all preparatory practices similarly; they may engage in some preparatory practices more frequently and some less so, and their preparatory practices also have similarities and differences among different tests. The following is a comparison of the three tests on five categories of preparatory activities.

The five types of preparation practices are test management, drilling, memorisation, language skill development and social affective strategies.

(1) In terms of test preparation management, on the whole, test-takers employ this practice with relatively low frequency and no significant difference is found.

(2) In terms of drilling, or mass practice, results show that *IELTS* and TOEFL iBT test-takers take up drilling significantly more frequently than CET-6 test-takers.

(3) In terms of memorization, the results show that test-takers employ memorisation practice with relatively high and similar frequency.

(4) In terms of language skill development, the results show that test-takers resort to language skill development with relatively high and similar frequency.

(5) In terms of social affective practice, the results show that in general, test-takers employ social affective practice with relatively low frequency, but *IELTS* and TOEFL iBT test-takers tend to take this practice significantly more often than CET-6 test-takers.

As for the time investment on preparation, most CET-6 respondents (60.4%) spend less than half of a month on preparing CET-6; however, the majority of the *IELTS* and TOEFL iBT test-takers spend over a month on preparation and averagely spent more than half an hour on preparing writing every day. The results suggest that *IELTS* and TOEFL iBT writing exerted longer washback on test-takers than CET-6 writing.

## 4 Conclusion

In conclusion, for the first research question, this study finds that there are similarities and differences concerning the washback of CET-6, *IELTS* and TOEFL iBT writing tests on test-takers' perceptions.

In general, *IELTS* and TOEFL iBT writing tests exert significantly stronger influences on test-takers' perceptions of test design, self-concept of their abilities and task expectancy than the CET-6 writing test.

The three writing tests exert similar influences on test-takers' perceptions of test uses and subjective task values.

For the second research question, this study finds that there are similarities and differences concerning the washback of CET-6, *IELTS* and TOEFL iBT writing tests on test-takers' preparation. In general, *IELTS* and TOEFL iBT writing tests impose more intense and longer washback on test-takers' preparation.

For preparation practice, *IELTS* and TOEFL iBT test-takers use drilling and social affective strategies significantly more frequently than CET-6 test-takers. However, test-takers of the three writing tests take up test preparation management, memorisation and language skill development with no significant difference.

For time investment, *IELTS* and TOEFL iBT test takers spend significantly longer time on preparation than CET-6 test takers.

In summary, this study comparatively investigates the washback of CET-6, *IELTS* and TOEFL iBT writing tests on test-takers in China, specifically focusing on the washback on test-takers' perceptions and test preparation processes. The study provides valuable insights into the perceptions and preparation practices of the influential group of stakeholders—test takers who are of particular relevance to the exams developers, researchers, and test users. It is hoped that this study merits further investigation into writing test washback on learners in the Chinese context and beyond. More importantly, it is intended to provoke reflective thoughts about how to positively and effectively promote learners' writing learning with the help of proficiency tests like CET-6, IELTS and TOEFL iBT.

## References

Hughes, A. (1993). *Washback and TOEFL 2000*. Unpublished manuscript, University of Reading.

Jacobs, J. E. & Eccles, J. S. (2000). Parents, task values, and real-life achievement-related choices. In C. Sansone & J. M. Harachiewicz (Eds.), *Intrinsic motivation* (pp. 405–439), San Diego: Academic Press.

Xie, Q. (2010). *Test design and use, preparation, and performance: A structural equation modeling study of consequential validity*. Unpublished PhD thesis, the University of Hong Kong.

## Funding

This study is part of the achievements of the Key Research Project of the National Philosophy and Social Science Foundation of China [Grant No.14AYY010].

# The Generation of an Individualized Cognitive Diagnostic Report for College English Writing

**Tan Yandan**, Xi'an Jiaotong University, China
**Ma Xiaomei**, Xi'an Jiaotong University, China
**Lu Chang**, Xi'an Jiaotong University, China

**Abstract:** Cognitive diagnostic assessment (CDA) is specifically designed to measure a student's knowledge structures and processing skills. This study was intended to develop a cognitive diagnostic assessment model for college English writing and to evaluate the feasibility of applying its diagnostic report. The Reduced RUM available in the software R studio was adopted as a major research tool to get the students' mastery profile of each attribute by inputting data of the Q matrix and test-takers' performance. This study followed the framework presented by Roberts & Gierl (2010) with the purpose of developing score reports for cognitive diagnostic writing assessments. The individualised cognitive diagnostic report provided learners with a total score, a specific attribute mastery profile, and strategies for improving the weaknesses, which would be beneficial to both students and teachers in English writing.

## 1 Introduction

The goal of college English teaching is to cultivate students' English competence, to enhance the intercultural communicative awareness and communicative competence, and to develop self-learning ability and improve their comprehensive cultural literacy. The ability to write effectively, as one of the vital parts of English competence, is becoming increasingly important in the global community, and instruction in writing is thus assuming an increasing role in both second and foreign language education (Weigle, 2002). However, ESL (English as a Second Language) writing has been widely recognised as the difficulty of second language learning and instruction in China, being regarded as the top priority of common concern. Moreover, responding to students' writing is an important aspect of second language (L2) writing programs that is fundamentally concerned with the successful development of their L2 writing skills.

As the role of writing in second language education increases, there is an ever greater demand for valid and reliable ways to test writing ability, both for classroom use and as a predictor of future professional or academic success. With the encouragement of "assessment for learning", diagnostic assessment is a subject of increasing interest in the language assessment community, as researchers, recognising the limitations of proficiency tests, have turned their attention to assessments that contribute to instruction and curriculum improvement (Alderson, 2005, 2007; Shohamy, 1992). Furthermore, cognitive diagnostic assessment (CDA) has also been the cause of significant advancements in diagnostic assessment, although it is a relatively new concept. CDA assumes that the latent ability is composed of a set of knowledge structures, skills, or attributes. Students' probability of mastering each attribute can be calculated, and then student skill profiles are formulated. The results of a CDA yield a profile with specific information about a student's cognitive strengths and weaknesses, which has the potential to guide instructors, parents, and students in their teaching and learning processes. Compared with conventional summative assessment which only focuses on ranking students according to a single total score, CDA is fundamentally different, and its advantage lies in its ability to provide fine-grained diagnosis and personalised guidance.

## 2 Literature review

The recent applications of CDA in language assessment is extremely confined to reading and listening skills (Jang 2009; Ravand 2015; Yi 2016), and only few researchers have focused on cognitive diagnostic writing assessment. Kim (2010, 2011) examined the extent to which the diagnostic information generated by the Reduced Reparameterized Unified Model was a discriminant, accurate, and reliable method of determining student performance in English for Academic Purposes (EAP) writing. 10 English as a second language (ESL) teachers assessed 480 TOEFL iBT independent essays using the Empirically-derived Descriptor-based Diagnostic (EDD) checklist, which consisted of 35 concrete, fine-grained descriptors. The resultant ratings were then analysed using Arpeggio, the estimation software of the Reduced RUM. The findings showed that the skills diagnosis approach not only classified skill masters and non-masters accurately and reliably, but that it also had high discriminant function, with only a small number of students classified into flat profiles. He also gave several suggestions for future research. First of all, greater incorporation of students' perspectives should be considered as it was scarce in current scale development literature. In addition, potential applications for CDA in the area of integrative assessment are of particular interest. The final recommendation is associated with the EDD checklist's use in real classroom teaching and learning settings. Teachers might want to use the checklist to track students' writing performance over time, so that students receive both short- and long-term feedback. This continued investigation would be particularly important. Xie (2017) utilised a fine-grained diagnostic checklist to assess first-year undergraduates in Hong Kong and evaluated its validity and usefulness for diagnosing academic writing in English. Ten English language instructors marked 472 academic essays with the checklist. They also agreed on a Q-matrix, which specified the relationships among the checklist items and five writing subskills. This conceptual Q-matrix was refined iteratively via fitting a psychometric model (i.e. the reduced reparameterised unified model) to empirical data (i.e. the checklist marks) through the computer program Arpeggio Suite. The final Q-matrix was found to be valid and useful; it had far fewer parameters but greater power to discriminate masters and non-masters of academic writing skills. She found that the cognitive diagnostic model (CDM)-based skill diagnosis could identify the strengths and weaknesses in the five writing subskills for students across three proficiency levels and could provide richer and finer information than the traditional raw score approach. Although there are only two empirical studies in cognitive diagnostic writing assessment, they provide profound insights and reference for the current study in terms of methodology. However, there are few empirical CDA studies in college English writing, and almost none of them has focused on individualised cognitive diagnostic writing reporting.

## 3 Purpose and significance

This study was intended to develop a cognitive diagnostic assessment model for college English writing and to evaluate the feasibility of applying the report generated by cognitive diagnostic writing assessment in college English writing. In order to fulfill the purposes of the present study, the following three questions will be addressed:

(1) How is the cognitive diagnostic assessment model for college English writing constructed?
(2) What kind of individualised cognitive diagnostic report of college English writing could be generated?
(3) How do college students perceive the cognitive diagnostic writing report?

By answering these three research questions, the practicality of the cognitive diagnostic writing assessment can be evaluated, which helps to find out the acceptability of the cognitive diagnostic writing report and whether a finer-grained diagnostic writing feedback appeals to students' needs. Besides, the construct validity of the cognitive diagnostic writing assessment would be inferred. Therefore, this study is of great significance in both theory and practice.

Theoretically speaking, this study can make a certain contribution to both fields of ESL writing and CDA. As for the ESL writing, this study will provide empirical evidence for the acceptability of cognitive diagnostic writing feedback. Moreover, this study could also set an example for providing students with more fine-grained feedback which is totally different from the traditional teacher feedback and peer feedback. Concerning the CDA, this study is the right response to the call of CDA feedback research, since it is a vital part of CDA research with less achievement. Furthermore, cognitive diagnostic language assessment has focused more on reading and listening skills instead of writing skills since it requires more time and energy. This study, however, is definitely one of the rare cases of cognitive diagnostic writing assessment research.

Practically speaking, the individualised cognitive diagnostic writing report can provide learners with a total score, a specific attribute mastery profile, and strategies for improving the weaknesses, which would be beneficial to both students in the respect of ESL writing. Students could be more clear-minded about their writing ability, especially their shortcomings in certain aspects. What's more, the diagnosis and improvement strategies provided by the cognitive diagnostic report would be significant for them to make their future study plans.

## 4 Methodology

This study used a mixed methods research design, comprising both quantitative and qualitative approaches in order to gather multiple sources of empirical evidence. To be more specific, English writing tests were adopted to gather quantifiable data, while questionnaires and semi-structured interviews were employed to collect qualitative data. 70 freshmen at Xi'an Jiao Tong University and 44 sophomores in Xi'an Jiao Tong University City College were the main participators in this study. Three experienced raters also took part in this study.

In order to gather quantifiable data from cognitive diagnostic writing assessment, we had to construct the cognitive diagnostic assessment model for college English writing. The first thing was to develop descriptor-based checklists which could be used to rate essays. This study mainly applied think aloud protocol to empirically generate the descriptors. There were 10 experienced English teachers involved in this session and the follow-up interview. Afterwards, all these descriptors were analysed and arranged, and the empirically derived descriptor-based (EDD) checklist (Lu, 2017) was settled. Secondly, we also had to define the attributes involved in college English writing. Through referring to previous literature, teaching and testing syllabus and expert judgement, all the five attributes in this study were identified and defined. The third step was to construct the Q-matrix which showed the relationship between descriptors and attributes. Each descriptor could be related to one or more than one attribute, and each attribute was involved in at least three descriptors.

Since all the required elements for cognitive diagnostic writing assessment were ready, we could conduct a writing test for 114 students from four English classes. Students were asked to write on three different College English Test (CET) band 4 prompts. When all the essays were rated by using the EDD checklist, we applied R studio to generate the students' skill mastery profiles, which were summarised in the individualised cognitive diagnostic writing report. After delivering the report to some students, they were asked to do a survey online to help us to know about their perceptions to the cognitive diagnostic writing report, and 10 of them were further interviewed to investigate their use of the report. There are three fixed eliciting questions in the interview, listed as follows:

(1) Have you ever referred to the diagnostic information in the cognitive diagnostic writing report in your daily study? If yes, how often do you refer to it? If no, why are you not taking a look at it?

(2) Which part of the report do you think is the most useful?

(3) Do you think the report has helped you in improving your English writing ability especially for your weakness? If yes, can you give some examples? If no, can you give some reasons?

## 5 Results

### 5.1 How is the cognitive diagnostic assessment model for college English writing constructed?

The CDA model for college English writing (Lu, 2017) was established through three major steps. First, develop and validate the descriptors by think aloud protocol and multi-faceted Rasch analysis. Second, identify the attributes by reviewing literature and syllabi. Third, construct and validate the Q-matrix by expert judgement and clinical data fit. When all these three steps were finished, the model was generated.

### 5.2 What kind of individualised cognitive diagnostic report of college English writing could be generated?

When the model was settled, the cognitive diagnostic report could be generated. Based on it, we developed an individualised diagnostic feedback which consists of 1) a total score, 2) a specific attribute mastery profile, where individual learners' strengths and weaknesses were displayed in verbal form following a bar chart, 3) a detailed rating result of the checklist as a reference and 4) strategies for improving their weaknesses.

### 5.3 How do college students perceive the cognitive diagnostic writing report?

We gave out the diagnostic report and collected all the data of the survey about students' perceptions towards the diagnostic report. The survey contained 11 statements like "The cognitive diagnostic writing report correctly reflects my English writing ability", which were related to students' views and attitudes on the content and format of the cognitive diagnostic writing report. The survey adopted the Likert scale, ranging from strongly disagree to strongly agree. 73% of the students agreed or strongly agreed with the cognitive diagnostic writing report; 18% of the students were uncertain; only 7% of the students disagreed; and 2% of the students strongly disagreed. The results indicate that students accept and recognise the cognitive diagnostic writing report.

To gather more detailed information, we interviewed 10 students. According to the results, all the students have carefully referred to the cognitive diagnostic writing report, especially the improvement strategies, but the frequency of reference was limited. Some students mentioned that they had adjusted their study plans based on the suggestions of the cognitive diagnostic writing report. Some transcripts are as follows.

R: Which part of the report do you think is the most helpful for you?
S1: The Improvement Strategy.
R: Why do you say that?
S1: Because it reflected the problems I had in the process of writing.
R: After receiving this report, did you have a study plan for English writing based on the feedback?
S4: I have adjusted my study plan according to this report.
R: Do you think this report can help you improve your English writing skills, especially on your weaknesses?
S4: Yes, I think so.
R: Can you give an example?
S4: The report pointed out that my words were not appropriately used. In fact, I do feel that my vocabulary is not enough, so recently I have started to remember new words.

*(R =  Researcher; S = Student)*

## 6 Conclusions

To summarise, the cognitive diagnostic writing assessment model was successfully constructed and proved to be workable for college English writing. The model could also pave a way for the advancement of a computerised cognitive diagnostic writing system. Moreover, the cognitive diagnostic writing report was discriminating and accurate, which was regarded as helpful for college students in improving English writing ability, since it has identified students' strengths and weaknesses and provided them with tailored feedback and suggestions. However, the effectiveness of the cognitive diagnostic writing report in learning and teaching deserves more attention and in-depth research. In general, this study has significant findings on "personalised learning" and "autonomous learning".

## References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The Interface between learning and assessment*. London: Continuum.

Alderson, J. C. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 21–39). Ottawa: University of Ottawa Press.

Jang, E. E. (2009). Cognitive diagnostic assessment of l2 reading comprehension ability: validity arguments for fusion model application to. *Language Testing, 26*(1), 31–73.

Kim, Y. H. (2010). *An argument-based validity inquiry into the empirically-derived descriptor-based diagnostic (EDD) assessment in ESL academic writing*. Unpublished doctoral dissertation, University of Toronto, Toronto.

Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing, 28(4)*, 509–541.

Lu, C. (2017). *Development and validation of a cognitive diagnostic model for college English writing*. Unpublished MA thesis, Xi'an Jiaotong University.

Ravand, H. (2015). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 782–799.

Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement Issues & Practice*, *29*(3), 25–38.

Shohamy, E. (1992). Beyond proficiency testing: a diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal, 76*(4), 513–521.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Xie, Q. (2017). Diagnosing university students' academic writing in English: is cognitive diagnostic modelling the way forward? *Educational Psychology, 37*(1), 1–22.

Yi, Y. S. (2016). Probing the relative importance of different attributes in l2 reading and listening comprehension items: an application of cognitive diagnostic models. *Language Testing*, *34*(3), 337–355.

## Funding

This is a study of the research project granted by China National Social Science Fund (Project No. 17BYY015).

# Assessment in the future: A Cognitive Diagnostic Modelling for College English Reading Test

**Du Wenbo**, School of Foreign Studies, Xi'an Jiaotong University, China
**Ma Xiaomei**, School of Foreign Studies, Xi'an Jiaotong University, China

**Abstract:** Diagnostic language assessment (DLA) has recently gained much attention from teachers, language testers and second language acquisition researchers. It seeks to promote further learning designed to address the learners' weaknesses and increase their overall growth potential with "learning-oriented assessment" as its rationale. With the empowerment of Cognitive Diagnostic Approach (CDA), DLA is possible to be achieved not only theoretically but also methodologically and practically. However, the results of most previous CDA-based research have classified learners' mastery of a set of tested skills into a dichotomous-score pattern (0/1 pattern), lacking accuracy in that the critical value of mastery is vague. Aiming at providing students with finer-grained diagnostic feedback, this study addresses how a College English reading diagnostic assessment model is constructed via the combination of CDA and tree-based regression (TBR). Group level and individual level diagnostic results were successfully generated and synthesised into carefully designed diagnostic feedback as a final outcome.

## 1 Introduction

Traditional EFL reading assessment has been criticised for lacking diagnostic information to inform students of their strengths and areas for improvement. With the aim to fill this gap, diagnostic language assessment (DLA) appears on the scene and has been arousing much interest in the field of second language assessment.

DLA is defined as the processes of identifying students' weaknesses, as well as their strengths, in a targeted domain of linguistic and communicative competence and providing specific diagnostic feedback and guidance for remedial learning (Lee, 2015). With the refinement of advanced psychometric techniques, cognitive diagnosis approach (CDA) has been widely used among researchers to fulfill the above purpose in the existing DLA methodology (Buck & Tatsuoka, 1998; Buck, Tatsuoka, & Kostin, 1997; Jang, 2005, 2009; Kim, 2015; Lee & Sawaki, 2009; Leighton & Gierl, 2007; Jang, Dunlop, Park, & Vander Boom, 2015, etc). However, the results of most CDA-based research have classified students' mastery of a set of tested skills into a dichotomous-score pattern (0/1 pattern), lacking accuracy in that the critical value of mastery is vague.

Aiming at providing students with finer-grained diagnostic feedback, this study manages to apply a new approach called tree-based regression (TBR) along with the CDA approach to construct a cognitive diagnostic model for an EFL reading test. In light of the final goal, the major question this study aims to answer is: How to use tree-based regression analysis to construct a cognitive diagnostic model for an EFL reading test?

Three sub-questions are then proposed to address the major question:

(1) What cognitive attributes are indispensible in English reading?
(2) What is the mastery status of each reading attribute at group level?
(3) What is the mastery status of each reading attribute at individual level?

## 2 Literature review

### 2.1 CDA approach

CDA is a newly developed approach which is aimed at providing formative diagnostic feedback through a fine-grained reporting of learners' skill mastery profiles in a specific discipline (Chipman, Nichols, & Brennan, 1995; DiBello, Roussos, & Stout, 2007; Embretson, 1998; Hartz, 2002; Tatsuoka, 1983). Three core elements in CDA need to be addressed: cognitive attributes, Q-matrix and cognitive diagnostic model (CDM).

Cognitive attributes refer to a series of cognitive skills, strategies, methods and knowledge that the learner might need to correctly complete a given task (Buck & Tatsuoka, 1998; Leighton & Gierl, 2007).

Q-matrix refers to a two-dimensional incidence matrix which reflects the relationship between cognitive attributes and test items (Tatsuoka, Birenbaum, & Arnold, 1989).

It is difficult to measure, diagnose, and assess individuals' inner psychological processing simply because we cannot directly observe the thinking process. What we can observe, however, is their responses to test items. Cognitive psychologists and scholars of psychometrics have made numerous efforts to develop a series of psychological measurement models with diagnostic function, generally referred to as CDMs, which have "evolved into powerful tools over the last 20 years, particularly into areas of educational and psychological measurement" (Rupp, Templin, & Henson, 2010). So far, there are more than 120 different CDMs, such as Tatsuoka's Rule Space model/methodology (Tatsuoka, 1983, 1989, 1995), Sheehan's Tree-based model (1997), Hartz et al.'s Reparameterized Unified Model or Fusion Model (2002), and Attribute Hierarchy Method (AHM) (Leighton, Gierl, & Hunka, 2004).

### 2.2 Previous CDA-based reading research

Different CDMs have been applied in previous CDA-based reading research. Jang (2009) investigated the validity of the CDA Fusion Model to an existing large-scale reading comprehension test. Wang & Gierl (2011) made diagnostic inferences about examinees' cognitive skills in critical reading, in which AHM was applied to a subset of SAT critical reading items and illustrated how this method can be used to promote cognitive diagnostic inferences. And more recently, Kim (2015) applied Fusion Model to diagnose 1982 test-takers' strengths and weaknesses in L2 reading of a placement test. However, some common limitations do exist.

First, the definition of cognitive attributes lack specificity, whose validation also lacks statistical evidence. Second, the instruments in most studies are non-diagnostic tests, whose diagnostic function remains questioned. Finally, the results, as stated earlier, lack accuracy to some degree. Therefore, research on EFL reading processes should be in pursuit of the following aspects: close to reality, ensure accuracy, and enhance practicality.

### 2.3 Tree-based regression (TBR)

Tree-based regression (TBR) is a kind of multivariate regression which can be defined as the study of a dependency relation between a goal variable and a set of independent variables. The final product of TBR is a model of this relationship (Luís, 1999). This model can be used either for understanding the interactions between the variables of the domain under study, or to predict the value of the goal variable of future instances of the same problem. Through TBR analysis, we can identify the cognitive processes required by each item in a test and explain the variability of the psychometric properties of the set of items in the test in terms of the cognitive processes identified (Enright, Morley, & Sheehan, 2002; Huff, 2003; Sheehan & Ginther, 2001). Detailed illustration can be seen in Sheehan (1997) and Gao & Rogers (2010). In the present study, TBR is mainly adopted to statistically validate the usefulness of the defined reading attributes.

## 3 Methodology

The study is conducted via three main stages, including Stage 1: Construction and validation of cognitive reading attributes; Stage 2: Construction of a CDM via TBR; Stage 3: Generating diagnostic feedback.

### 3.1 Participants

Coders: Eight content experts were invited as coders to participate in the process of determining cognitive reading attributes for this study. All 8 experts are EFL teachers with at least 15 years' teaching experience in university. They had considerable knowledge in EFL reading and language testing.

Test-takers: 12 non-English majors from Xi'an Jiaotong University (XJTU) volunteered for the think aloud section of this study. 740 freshmen from different majors in XJTU participated in the diagnosis section. They were organised to take the diagnostic reading test online, and their responses to each item were recorded in the database corresponding to their student number.

### 3.2 Instruments

A diagnostic reading test designed by PELDiaG[7] research team in XJTU was adopted. Informed by previous research i.e., the Dialang program (Alderson, 2005) and relevant reading testing theories, the test construct was mainly based on the College English Curriculum for reading and the CET-4 syllabus. The test paper was retrofitted following TOEFL reading, CET-4 reading, and self-design principles according to CDA procedure, including three sections with topics covering college life, social culture, science and literature and was designed to be completed in 90 minutes.

[7] PELDiaG refers to the web system named "Personalized English Learning: Diagnosis and Guidance" developed by the research team at Xi'an Jiaotong Unversity (XJTU).

### 3.3 Research procedure

In Stage 1, a set of cognitive reading attributes along with a hypothesised Q-matrix and a weighted Q-matrix are obtained. The attributes and Q-matrix are used to conduct TBR analysis in Stage 2, resulting in a tree-based model describing the relationship between item difficulty and reading attributes. The tree model is then translated into a group-level mastery tendency model by summarizing test-takers' performance on each identified reading attribute using a LOWESS approach, thus forming group-level diagnostic results. Finally, in Stage 3, individual level diagnostic information is presented in well-designed diagnostic feedback as an outcome of this study.

## 4 Results

### 4.1 What cognitive attributes are indispensible in English reading?

By reviewing reading theories and previous research upon EFL reading, the theoretical framework of cognitive reading attributes is first determined, i.e., language knowledge and strategic competence from the linguistic level, and surface code, test base, situational mode from the cognitive level. Subsequently, nine initial reading attributes are defined. After a two-round judgment among eight reading experts and modification of each reading attribute, eight hypothesised cognitive reading attributes are finally defined. They are *A1, understanding sentence literal meaning*; *A2, understanding discourse literal meaning* (two language knowledge attributes); *A3, seducing word meaning*; *A4, contextual inference*; *A5, elaborative inference*, *A6, synthesising and summarising*; *A7, locating relevant information;* and *A8, eliminating alternative choices* (five strategic competence attributes). Students' think aloud protocol testifies the validity of the defined reading attributes qualitatively, and the attributes are then used to construct a Q-matrix. Seven reading experts code the data, and their internal agreement value calculated using Fleiss Kappa is 0.41, reaching a moderate agreement among coders. A hypothesised Q-matrix and a weighted Q-matrix are obtained subsequently. The weighted Q-matrix and defined reading attributes are complied to run TBR analysis. A tree-based model describing the relationship between item difficulty and cognitive reading attributes are successfully generated. The TBR results show that the eight reading attributes account for 73.63% of the variance in item difficulty, thus validating the meaningfulness of the reading attributes from a quantitative aspect, providing strong evidence of the effectiveness and practicality of the defined attributes. Eventually, eight cognitive reading attributes are recognised as indispensable in EFL reading and validated quantitatively and qualitatively.

### 4.2 What is the mastery status of each reading attribute at group level?

As for group level mastery status, the resulting eight LOWESS curves are all in an increasing trend and bound at one, which vividly depict the tendency of group attribute mastery probability. Specifically, the overall group mastery probabilities show that the two language knowledge attributes are mastered best in the group, while A5, elaborative inference; and A3, deducing word meaning, have the two lowest mastery probabilities. The mastery probability of the other four attributes (A4, A6, A7, and A8) range from 70.64% to 72.63%, showing a low variability. Therefore, intervention and remedial solutions should be made to help students with their skill areas for improvement, especially on A5 and A3. Finally, the group level diagnostic result is then provided to students, teachers and teaching administrators in a bar chart format.

### 4.3 What is the mastery status of each reading attribute at Individual-level?

We selected 15 students with the same total score (72) to analyze any individual differences in attribute mastery status. The results showed that individual students' attribute mastery probabilities vary from person to person. Students with the same total score demonstrate completely different mastery status of each cognitive reading attribute as a result of individual difference.

Finally, individual-level diagnostic feedback is shown in a designed form, from which students could directly see what their strengths and areas for improvement. Compared to the result of traditional reading tests, which generate superficial total score, the customised diagnostic feedback developed in this study has the following two advantages. On the one hand, it fulfills the ultimate diagnostic purpose required in DLA by providing much more detailed diagnostic information in addition to the total score. Each student taking part in the test will receive diagnostic feedback that includes their proficiency level relative to a particular group, their mastery profiles for each reading attribute, from which they can recognise their strengths and areas for improvement, and a detailed illustration and guidance with their future learning. Since individual differences do exist among students, such a customised diagnostic feedback is very likely to help students better understand their reading proficiency from different perspectives, thus stimulating and contributing to their remedial learning.

## 5 Conclusion

The purpose of this study is to apply a new method called tree-based Regression along with CDA approach to construct a CDM of EFL reading test with the ultimate goal of generating meaningful diagnostic feedback for both students and teachers to promote the learning of English reading. The study successfully generated both group level and individual level diagnostic information and presented to students, teachers and teaching administrators in a proper form as meaningful instructions. Despite instructive results, several limitations cannot be ignored.

Firstly, the participants in this study were 740 freshmen from Xi'an Jiaotong University. The size of the sample compared to other DLA research (Jang, 2009; Kim, 2015; Sheehan, 1997) is not big enough, which may influence the accuracy of some statistical results. And since all participants come from one university, the representativeness of this study still needs to be considered. The results might be different when it comes to students in other universities.

Secondly, though efforts and modifications have been made on the eight identified reading attributes, they cannot cover all the cognitive processes or strategies in reading comprehension. Whether these eight reading attributes could also produce satisfactory results when used in other diagnostic reading materials still needs further validation. And the Q-matrix coded by 7 content experts reached a moderate agreement level due to individual difference among experts. Though it is acceptable, a substantial or perfect agreement is expected in the future research. A high internal consistence among experts may contribute to a clearer definition of reading attributes and the design of reading items.

Thirdly, it cannot be ignored that the Q-matrix developed by content experts may not fully reflect the actual attribute combinations students adopted when answering the reading questions. It is highly possible that students might refer to other cues to make correct response to an item. Therefore, defining and validating cognitive reading attributes, and Q-matrix, the crucial and necessary step of all DLA research, still have much room to improve.

## References

Alderson, J.C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. New York: Continuum.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test*. Language Testing, 15*(2), 119–157.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple choice test of second language reading comprehension. *Language Learning, 47*(3), 423–466.

Chipman, S. F., Nichols, P. D., & Brennan, R. L. (1995). Introduction. In S. F. Chipman, P. D. Nichols & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 1–18). Hillsdale: Lawrence Erlbaum.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics volume 26: Psychometrics* (pp. 970–1030). Amsterdam: North-Holland Publications.

Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 6*(3), 380–396.

Enright, M.K., Morley, M, & Sheehan, K.M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15*, 49–74.

Gao, L. & Rogers, W. T. (2010). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing, 28*(2), 1–28.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Huff, K. (2003). *An item modeling approach to providing descriptive score reports*. Unpublished doctoral dissertation, University of Massachusetts Amherst.

Jang, E.E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Jang, E.E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing, 26*(1), 31–73.

Jang, E.E., Dunlop, M., Park, G., & Vander Boom, E. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback?. Language Testing, 32 (3), 359–383.

Kim, A. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258.

Lee, Y.W. (2015). Diagnosing diagnostic language assessment. *Language Testing, 32*, 299–316.

Lee, Y. W. & Sawaki, Y. (2009). Cognitive diagnostic approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172–189.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka’s rule-space approach. *Journal of Educational Measurement, 41*(3), 205–237.

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York: Cambridge University Press.

Luís,F. (1999). *Inductive Learning of Tree-based regression models*. Unpublished doctoral dissertation, University of Porto.

Rupp, A.A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guildford Press.

Sheehan, K. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333–352.

Sheehan, K. & Ginther, A. (2001). *What do multiple choice verbal reasoning items really mea-sure? An analysis of the cognitive skills underlying performance on a standardized test of reading comprehension skill*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconception based on item response theory. *Journal of Education Statistic, 20*(4), 345–354.

Tatsuoka, K.K., Birenbaum, M., & Arnold, J. (1989). On the Stability of Students' Rules of Operation for Solving Arithmetic Problems. *Journal of Educational Measurement, 26*(4), 351-361.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statis¬tical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale: Law¬rence Erlbaum.

Wang, C. & Gierl, M. J. (2011). Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement, 48*(2), 165–187.

# The Assessment Literacy of Language Teachers: A Case Study of Teachers of Portuguese as A Foreign Language

**Maria Jose dos Reis Grosso**, University of Macao, China
**Catarina Isabel Sousa Gaspar**, Universidade de Lisboa, Portugal

**Abstract:** We propose an analysis of ideas and representations that teachers of Portuguese as a foreign language (PFL) have about assessment. Bearing in mind the existence of conceptual differences in curricula, contexts, learners and teachers' previous knowledge, we present the results of a survey that focuses on topics such as understanding what is considered of major importance in assessment, what is evaluated and how assessment is done when PFL teachers run a PFL course.

The ideas and representations of language teachers determine how assessment affects learners by a transfer process that influences the way they engage in the assessment of linguistic competence. We also discuss the need for developing flexible attitudes towards and personal confidence in assessment in the teaching and learning process.

## 1 Introduction

We propose an analysis of the ideas and representations of assessment that teachers of Portuguese as a foreign language (PFL) have. We present the results of a survey applied to teachers of PFL. The survey aimed to:

- clarify what is of major importance in assessment from the teachers' point of view
- know what they evaluate
- determine how they engage in assessment when they are running a PFL course
- discern whether they have received specific training in assessment
- evaluate the level of PFL teachers' language assessment literacy; and
- identify training needs.

The survey was intended to be just the beginning of a larger follow-up study on the creation of research tools to collect data on the Assessment Literacy of Portuguese as a Foreign Language teachers (ALPFL).

## 2 Framework

According to Fulcher (2012), the American Federation of Teachers Standards for Teacher Competence in Educational Assessment of Students (1990) was the earliest attempt to define assessment literacy. This document recognised the major importance of assessment and teachers' roles both inside and outside of the classroom:

> The scope of a teacher's professional role and responsibilities for student assessment may be described in terms of the following activities. These activities imply that teachers need competence in student assessment and sufficient time and resources to complete them in a professional manner.

The same document also stresses the strategic importance of developing professional skills in assessment during teacher training. "Assessment literacy" is a term coined by Stiggins (1991) to refer to the range of skills and knowledge that stakeholders need to deal with assessment. Later, Stiggins (1995) proposed what he called “Assessment Literacy Redefined”, which Inbar-Lourie (2008) summarised as:

> the ability to understand, analyse, and apply information on student performance to improve instruction . . . . Becoming assessment literate requires the attainment of “a toolbox” of competencies, some practical and some theoretical, on why, when and how to go about constructing a variety of assessment procedures (p. 389).

Assessment literacy is related to the idea of assessment as something that can be culturally different and dynamic. Several studies of assessment literacy have used surveys that gathered data about teachers’ real assessment competencies and literacy and their training needs, of which those of Fulcher (2012), Montee, Bach, Donovan, & Thompson (2013), Malone (2013) and Jannati (2015) are most notable.

Many of these studies have focused on teacher training and the dynamics of teachers’ practices inside the classroom. They emphasised that after years of research on assessment and the improvement of training programs, little has changed in classroom practices, especially in how teachers undertake assessment. Fulcher (2012) concluded: “Language teachers are very much aware of a variety of assessment needs that are not currently catered for in existing materials designed to improve assessment literacy” (p. 113).

Fulcher (2012) defined assessment literacy as follows:

> The knowledge, skills and abilities required to design, develop, maintain or evaluate, large-scale standardized and/or classroom based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice (p. 113).

Malone (2013) stated that “language assessment literacy refers to stakeholders (often with a focus on instructors’) familiarity with measurement practices and the application of this knowledge to classroom practices in general and specifically to issues of assessing language (Inbar-Lourie, 2008; Stiggins, 2001; Taylor, 2009)” (p. 330).

For PFL, we share the authors’ proposal that data gathered from studies/surveys on teachers’ (and other stakeholders’) assessment literacy can be used to create and improve assessment literacy resources, and to identify trends and practices inside classrooms that show teachers the necessity of training. As Malone (2013) states: “even when language instructors have received solid preparation about assessment during pre-service training, there is a need to provide on-going professional development to in-service teachers (Malone, 2008) as the field of language testing is in constant state of flux” (p. 332).

Changes have occurred in foreign language teacher training, especially with the spread of the Common European Framework of Reference for Languages (CEFR) effect and the international evaluation of educational systems under the patronage of European institutions. Language testing has undergone a process of growing professionalisation, which has led to an increase in language testing textbooks created by testing experts who may lack contact with the field (Davies, 2008; Malone, 2013) or with teachers and other stakeholders. In our study, we thus tried to avoid what Davies (2008) called “professional insularity”.

Bearing in mind this development of teaching and learning in PFL and the growth of the Centro de Avaliação de Português Língua Estrangeira (CAPLE) assessment network, we decided to focus on the assessment literacy of teachers engaged in teaching and learning PFL and on the application of CAPLE exams.

## 3 Methodology

An experimental survey was built in Google Forms and distributed by e-mail to the Centres of Administration and Promotion of Exams of Portuguese as Foreign Language (LAPE) in CAPLE's network. The person in charge of each centre was asked to share the survey with the PFL teachers with whom they worked. The total number of informants was 68 teachers, but between 66 and 68 responses to the questions were received.

Our net survey data are presented in graphs. Thus, the quantitative method is favoured, and as an exploratory study the qualitative method is subsidiary.

We adapted and used some of the survey questions created by Montee et al. (2013) that used a Likert-scale and were related to familiarity with assessment terminology (ALTE, 1998) and to the importance and frequency of assessment practices. To collect specific data, we added questions to determine the main demographic profile (age, sex, country and mother tongue) of the participants. Our survey also asked about professional experience, academic profile and whether the respondents had ever had any specific training in assessment and evaluation practices. Finally, the teachers were asked about the success of their students.

## 4 Data analysis

### 4.1 Gender

The respondents were mostly female (70.6% female and 29.4% male). This is an unsurprising finding in the area of education.

### 4.2 Age

The age of the respondents ranged from 22 to 74 years. This was divided into two large age groups, from 22 to 39 (48%) and from 40 to 74 (52%), respectively.

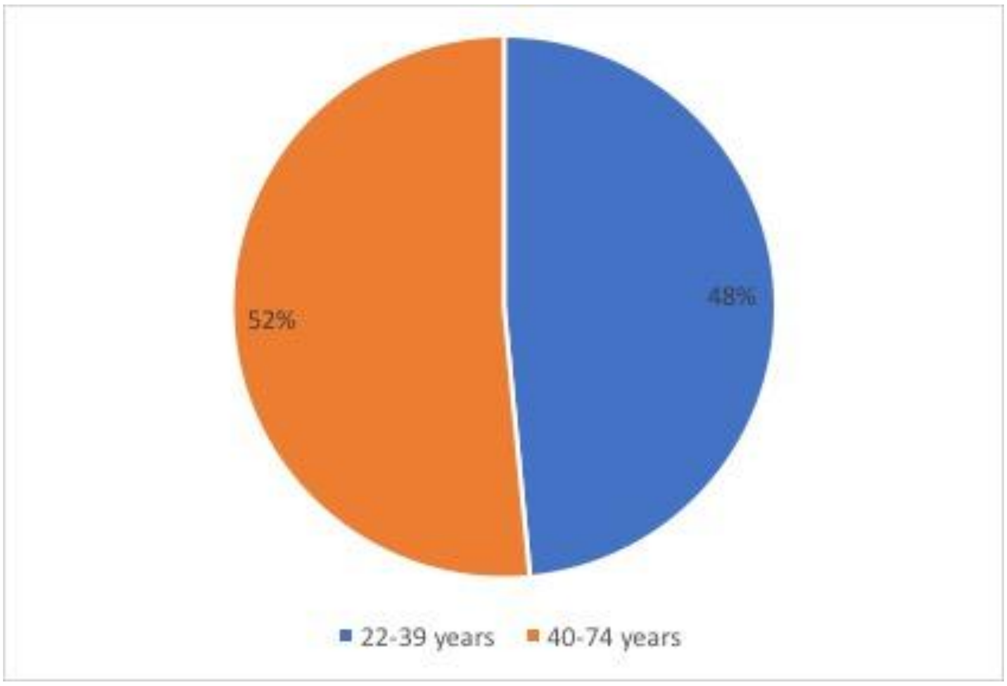


**Figure 1**. Age of respondents

### 4.3 Mother tongue: Portuguese

In terms of mother tongue, the teachers who taught PFL were mostly native speakers of Portuguese (91%), and only a small group did not have Portuguese as their mother tongue (they were Spanish and Mandarin speakers).

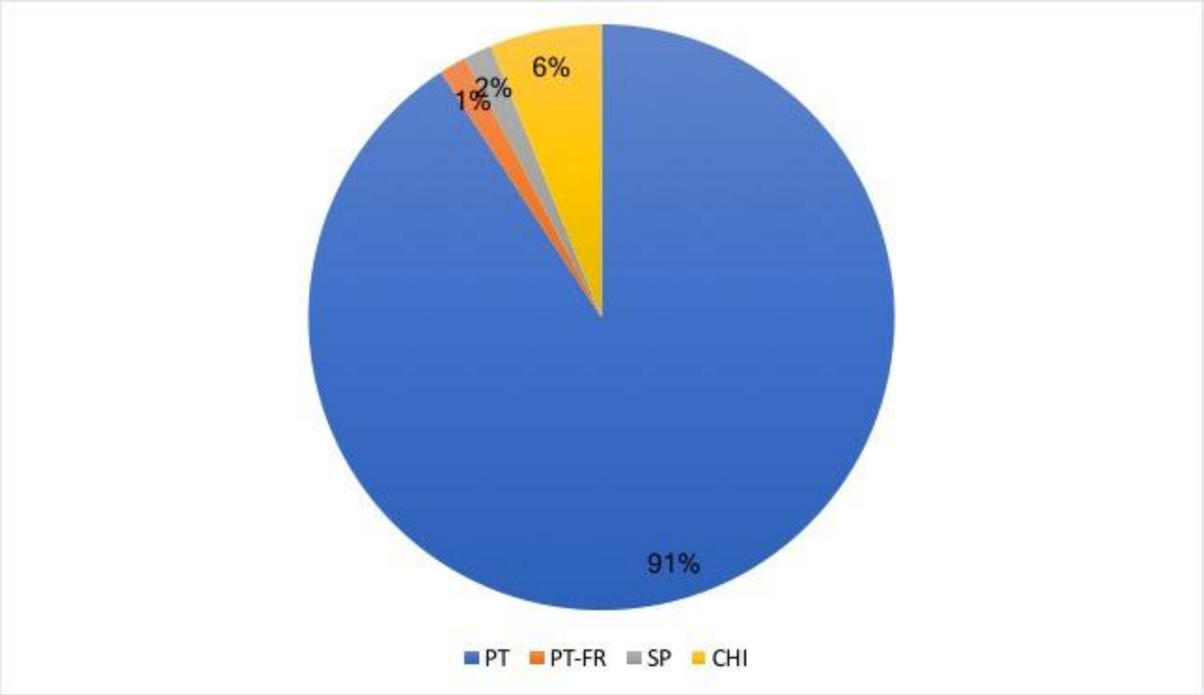


**Figure 2**. Mother tongue of respondents

### 4.4 Countries

Of the 96 exam application centres to which the survey was sent, we received answers from teachers from the following places: Andorra, Australia, China, Croatia, Egypt, Spain, France, India, Italy, Japan, Macau, Mexico, Portugal, Czech Republic, Rwanda and Switzerland.

| Country | Answers |
|---|---|
| Andorra | 1 |
| Australia | 1 |
| Czech Republic | 1 |
| China, Mainland | 4 |
| China, Macao | 12 |
| Croatia | 1 |
| Egypt | 1 |
| France | 1 |
| Greece | 1 |
| India, Goa | 1 |
| Italy | 3 |
| Japan | 1 |
| Mexico | 2 |
| Portugal | 2 |
| Romania | 1 |
| Ruanda | 1 |
| Spain | 21 |
| Switzerland | 11 |
| No answer | 1 |

**Table 1**. Countries where respondents work

### 4.5 Academic profile and training

As regards the respondents' academic profile and training, 38% had a bachelor's degree, 34% had a Master's degree, 25% had a PhD and 3% had other degrees.

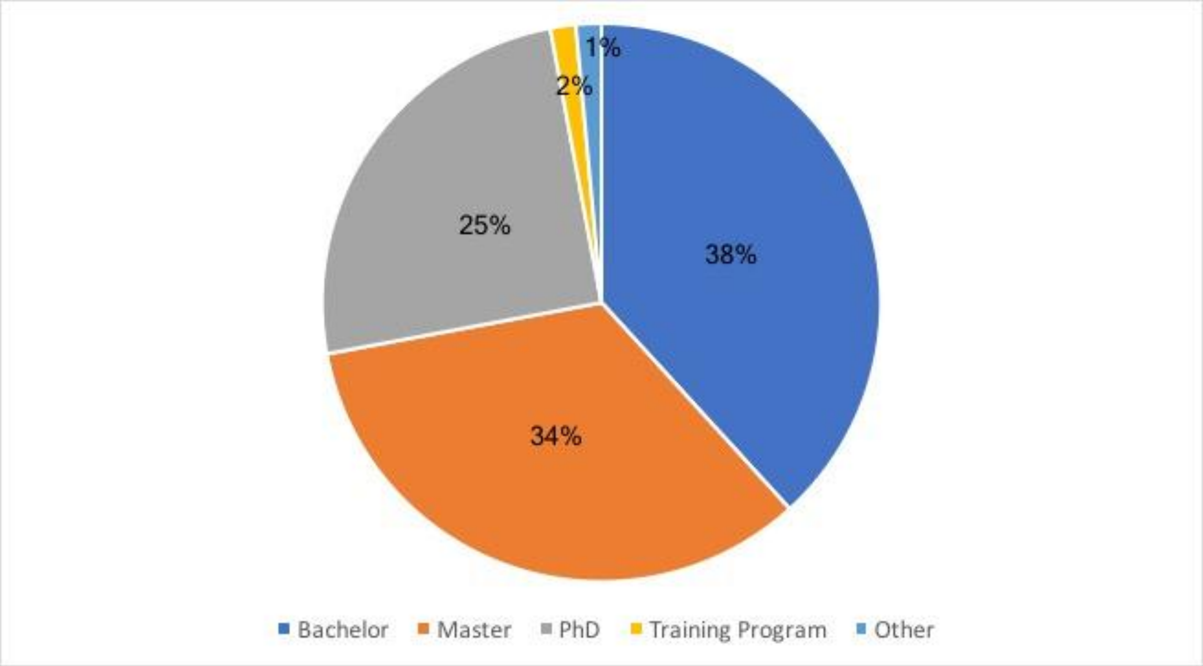


**Figure 3**. Academic profile of respondents

### 4.6 Working time: Full time

As can be seen from the following chart, most of the teachers who answered this questionnaire worked full time (79%), while only 21% said they worked part time.

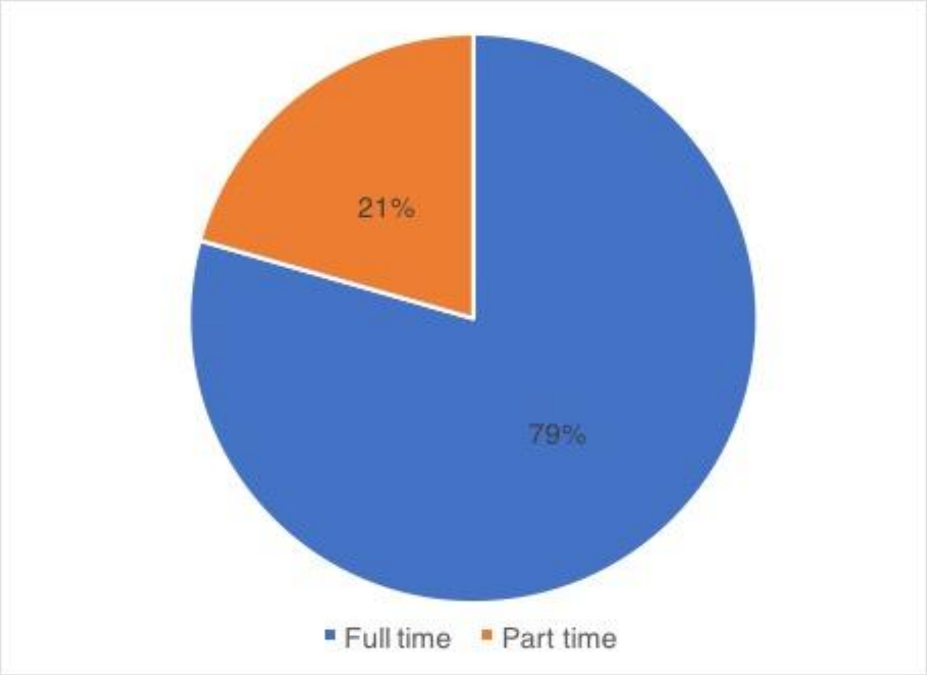


**Figure 4**. Working background of respondents

### 4.7 Teaching hours per week

Teaching time per week for the respondents ranged from those who did not teach at all to those with 30 hours of classes. The group with up to 20 hours was larger than the group with more than 20 hours.

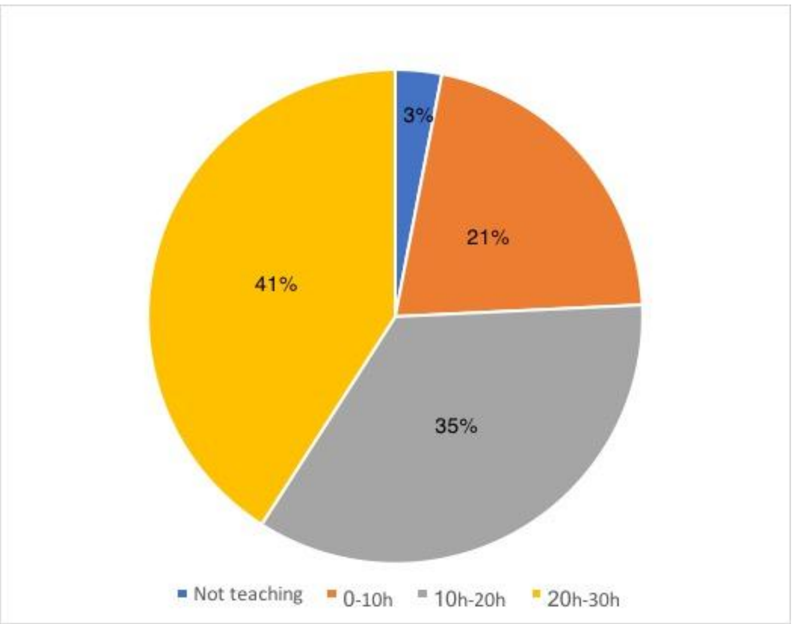


**Figure 5**. Teaching hours of respondents

### 4.8 Professional experience and time as a PFL teacher

In the area of time served as a PFL and professional experience, 45% of the teachers had worked for over 10 years and 13% had worked for only one year.

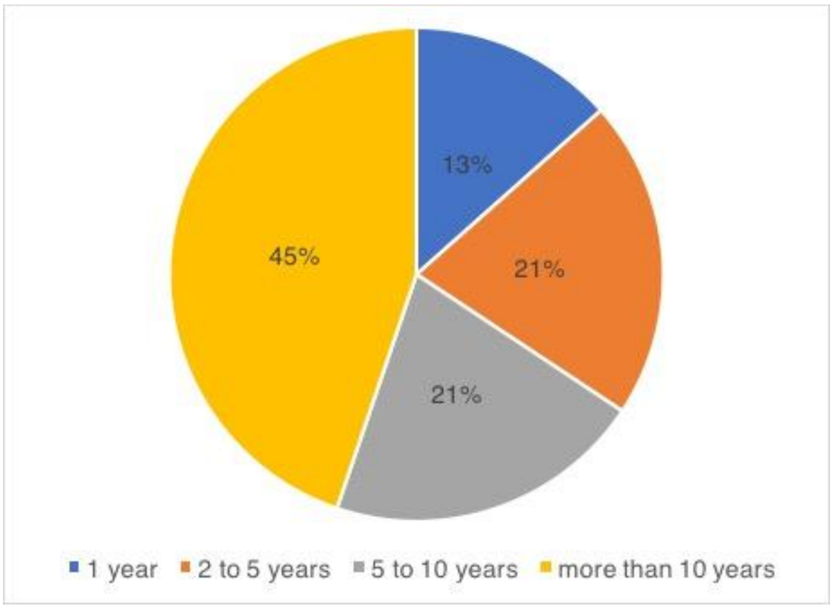


**Figure 6**. Professional experience of respondents

### 4.9 Specific training in assessment

79% of the teachers had never attended specific training in assessment, and only 21% answered affirmatively to this question.

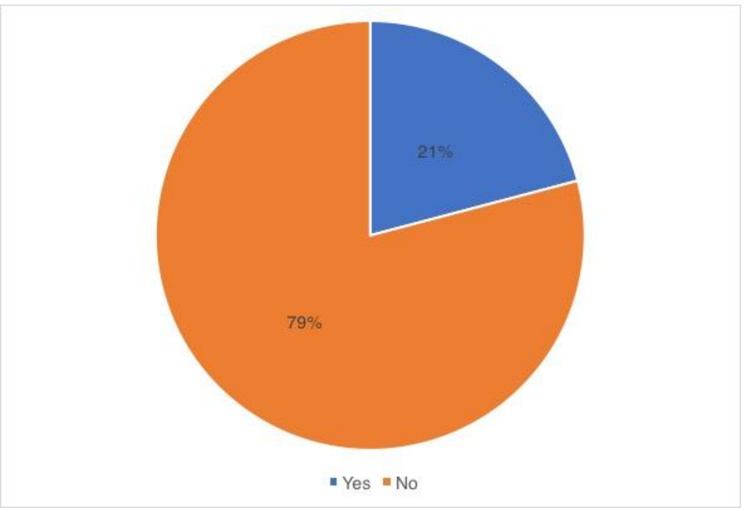


**Figure 7**. Number of respondents who had specific training in assessment

### 4.10 Prepared for assessment?

While a high percentage of the teachers had never attended a training evaluation, it is interesting to note that this did not modify their beliefs about being prepared to evaluate different activities. The respondents highlighted a readiness to evaluate reading comprehension and writing and to create tests or exams. The percentage answering ‘do not know’ and ‘little prepared’ was not very significant. The ‘little prepared’ response was mostly related to self-assessment, and assessing oral comprehension and speaking.

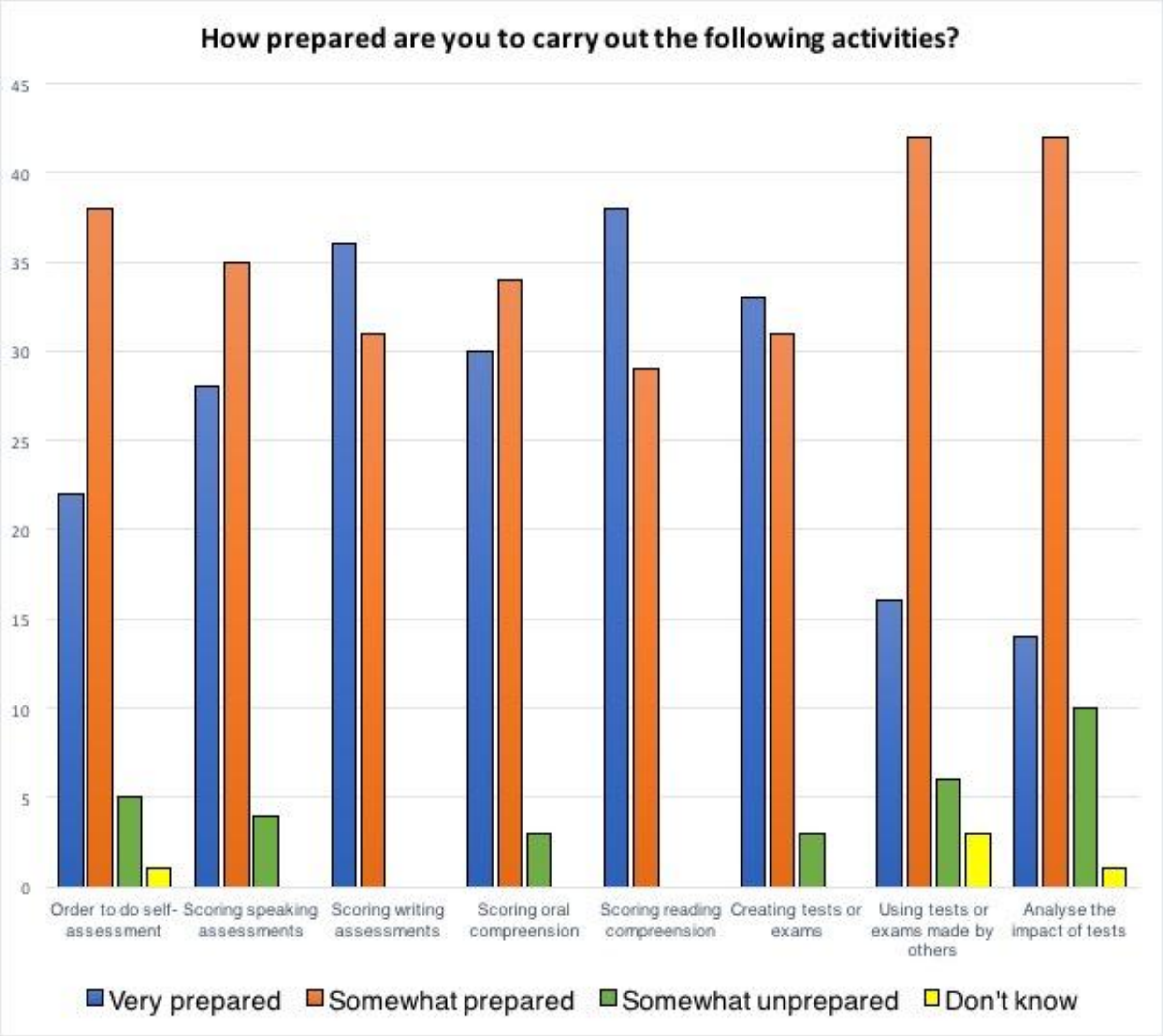


**Figure 8**. Respodents' level of preparation

### 4.11 Knowledge of terms related to evaluation and certification

Looking at the knowledge of and level of familiarity with terms related to the evaluation and certification of languages, the respondents mostly reported that they were well informed, responding with "I know it and use it". The graphs show that the teachers knew and used the terms "formative assessment", "summative assessment", "diagnostic test", "performance evaluation" and "tasks". The percentage of those who did not know these terms was small. Portfolios appeared to be little used as an evaluation tool; although the respondents knew of it, they did not use it.

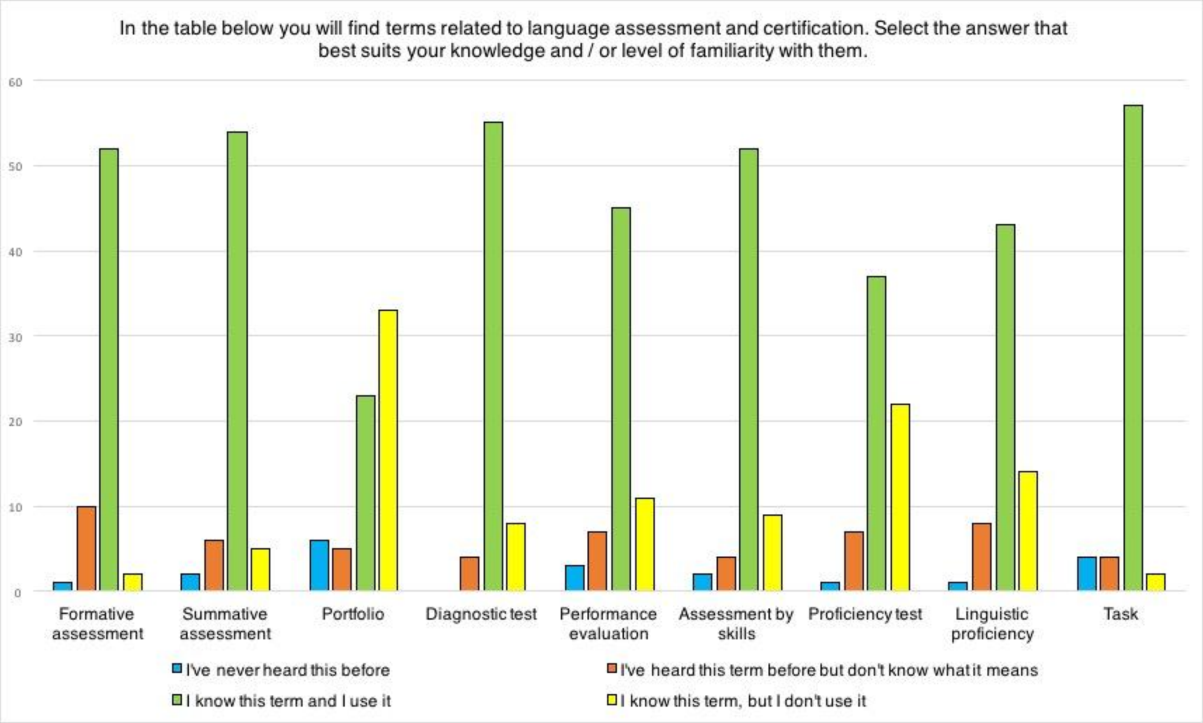


**Figure 9**. Respondents' knowledge of key terms (Part 1)

The graph in Figure 10 shows that most of the teachers knew all 14 of the terms shown. Terms such as 'sample', 'authenticity', 'holistic assessment', 'item bank' and 'classification' were less well known, although the terms were still used and understood by a significant number of teachers. The terms that were most unknown were 'calibration' and 'internal consistency'. Although virtually all of the terms were known, the following graph indicates that not everyone understood the terms.

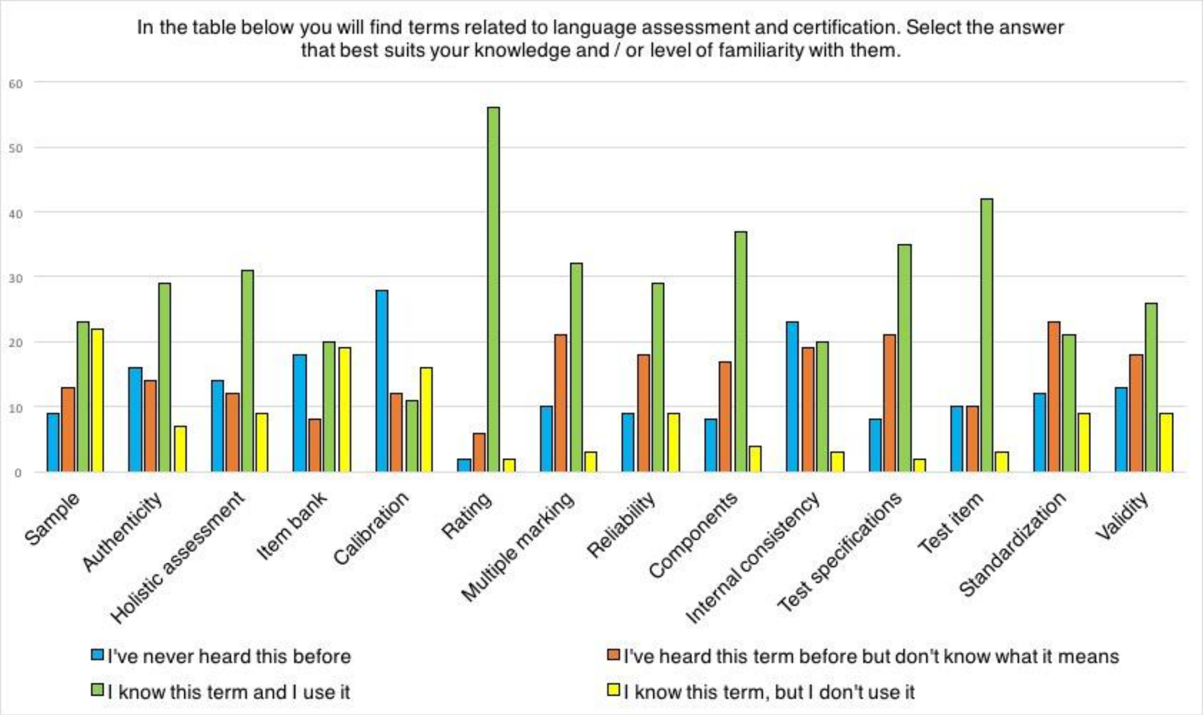


**Figure 10**. Respondents' knowledge of key terms (Part 2)

### 4.12 Relevance of assessing different skills

Almost all of the teachers stated that they attached great importance to oral and written comprehension and production skills, and it is interesting to note that the importance of orality was emphasised. However, reading comprehension was placed above written production.

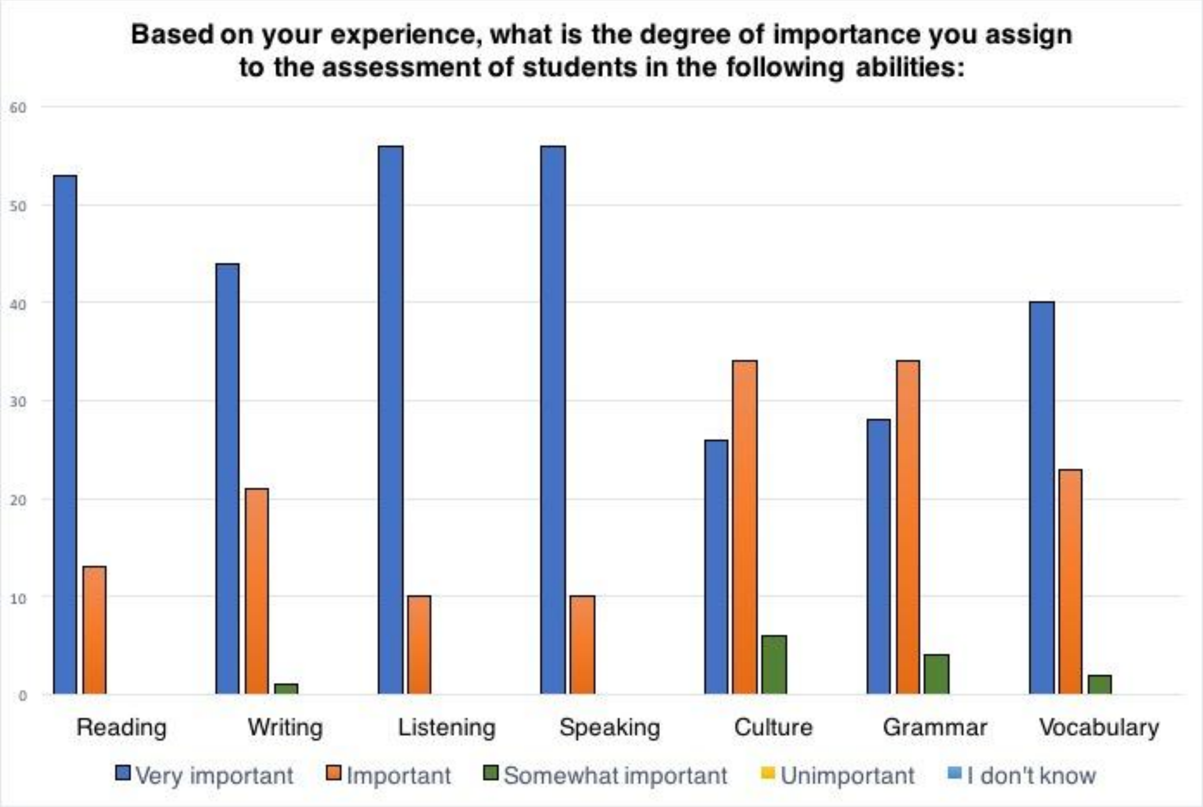


**Figure 11**. Respodents' rating of the importance of assessment topics

### 4.13 Frequency: How often are students assessed?

According to the data, the daily evaluation of reading comprehension and speaking was preferred. Writing was most commonly evaluated weekly, followed by monthly and then evaluation every two weeks. It is interesting to note that some of the respondents never evaluated speaking or oral comprehension, although the percentage was small. Across the skill areas, the proportion of semester evaluation was almost nil.

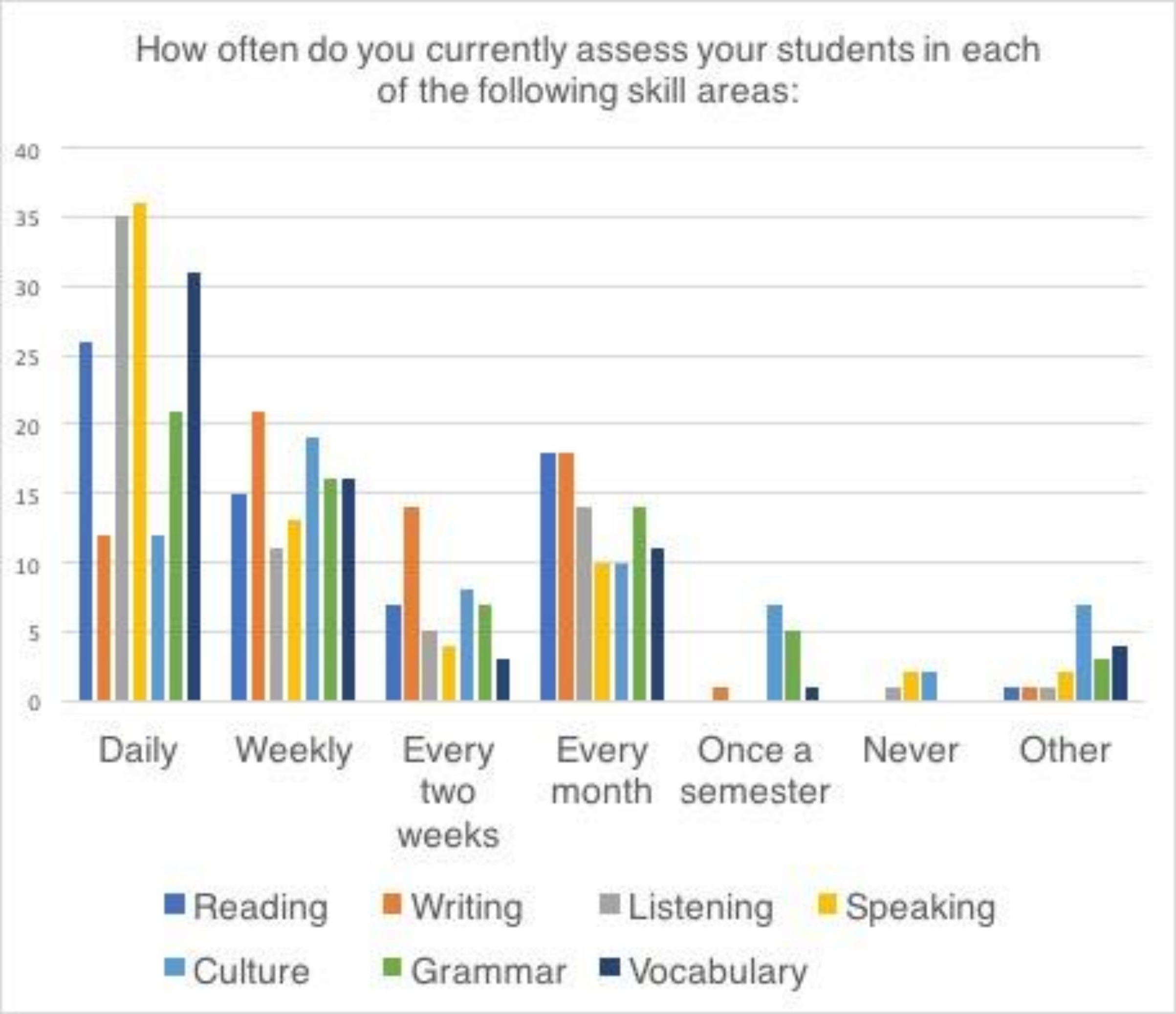


**Figure 12**. Frequency of student assessment by topic

### 4.14 Degree of satisfaction with the assessment performed

Over 70% of the respondents were satisfied with the assessment they made of their students. About 78% thought that students understood why they were being assessed, which indicates that the students understood how they were being assessed. About 61% of the respondents agreed that students should carry out self-assessment, but about 21% disagreed with self-assessment. Peer evaluation was not well accepted among 47% of the teachers; only 32% were in favour of this type of evaluation. The teachers were almost unanimous in knowing how to communicate the assessment to their students. About 91% used assessment results to plan their classes.

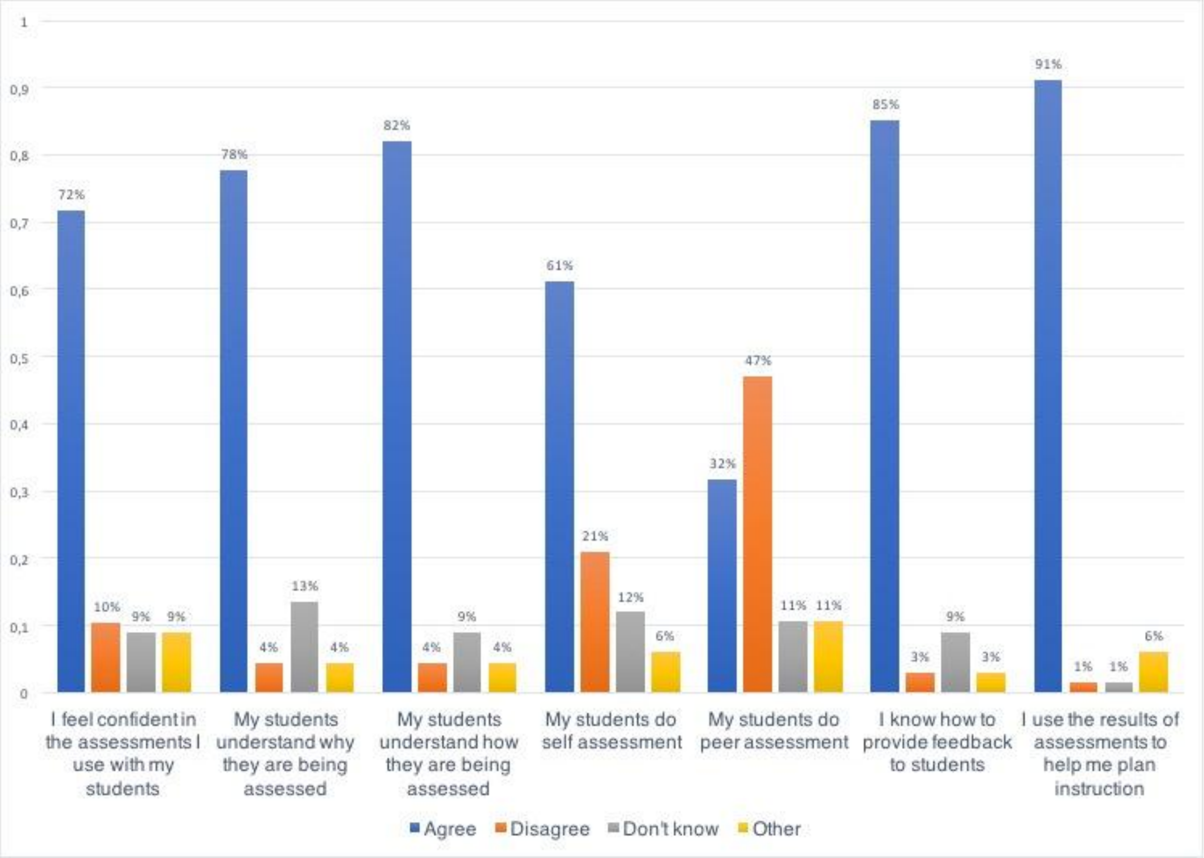


**Figure 13**. Respondents' confidence in their assessments

### 4.15 Positive results

As can be seen in the following graph, the main judgement category selected in assessment was positive/very good; about 91% of the respondents stated that more than 50% of their students succeeded. Notably, no-one chose the option of 0-10%.

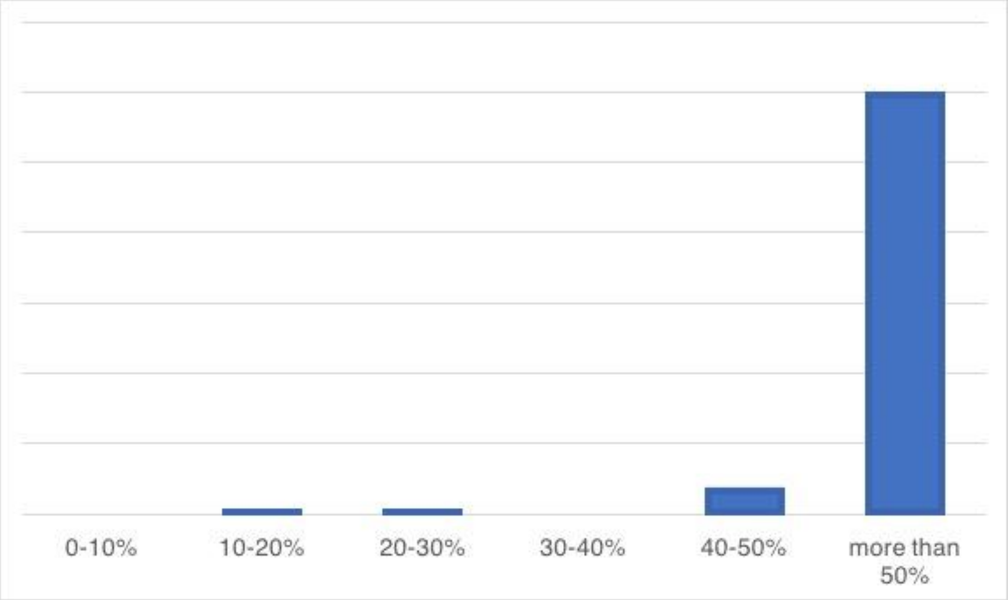


**Figure 14**. Respondents' evaluation of their students' success levels

## 5 Conclusions

It is interesting that only a small percentage (20%) of the respondents had specific training in evaluation, and this training, with few exceptions (from the pre-service education stage), was acquired during extracurricular activities.

In PFL, teachers' competence to assess usually emerges from informal on-the-job experience; it is rarely part of an explicit knowledge, much less formally conveyed.

The respondents were mostly teachers with professional experience and, according to their responses, were aware of the assessments they were making and were reasonably prepared to evaluate different language activities, comprehension and production skills, both oral and written.

Some 60% of the teachers applied self-assessment in their classes, but more than 20% felt less prepared to ask students to do self-assessments. Linked to this issue is that the teachers knew about language portfolios, but did not use them. In fact, although language portfolios are widely disseminated, teachers in remote contexts such as Macao do not use them much, nor do they attach great importance to them.

Knowing how to assess languages involves knowledge and the performance of various types of assessment that should, at the very least, contribute to improving the quality of students' learning through quality feedback (given by the teacher/evaluator).

Nevertheless, depending on context and educational tradition, there is a culture of summative and selective evaluation, as even formative evaluations are carried out with a summative purpose, and not to meet the needs of the student.

In this study, the teachers felt less prepared in assessing oral comprehension, or in preparing tests and exams administered one by themselves or others. They also felt less prepared to evaluate the impact of assessment on society. However, these results did not influence the PFL teachers' satisfaction with their evaluation endeavours: 70% were happy with the assessment of their students.

Assessment still seems to be centred on the teacher, as 46% of the respondents did not agree with peer evaluation. Nevertheless, teachers are informed by formal (or non-formal and informal) training in the learner-centred approach and understand the needs of target groups as being paramount in language teaching: in this study, 91% used assessment results to plan their classes, and the success rates for students were quite high.

We conclude that the importance of assessment must be reviewed and renewed; it must be more present in textbooks and in teacher training. The impact it causes on the life of the student and in society must be made more explicit. More importantly, language assessment features may be based on the informal knowledge of many experienced PFL teachers, a knowledge that must be made explicit through carefully crafted teacher-centred training.

## References

ALTE. (1998). *Multilingual Glossary of Language Testing*. Cambridge: UCLES/Cambridge University Press.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for Teacher Competence in Educational Assessment of Students*. Retrieved from http://buros.org/standards-teacher-competence-educational-assessment-students

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25* (3), 327–347.

Fulcher, G. (2012). Assessment literacy for the language classroom, *Language Assessment Quarterly, 9*, 113–132.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385–402.

Jannati, S. (2015). ELT Teachers' language assessment literacy: Perceptions and Practices. *Educational Research Association: The International Journal of Research in Teacher Education, 6*(2), 26–37.

Malone, M. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344.

Montee, M., Bach, A., Donovan, A., & Thompson, L. (2013). LCTL teachers' assessment knowledge as practices: An exploratory study. *Journal of the National Council of Less Commonly Taught Languages, 13*, 1–31.

Stiggins, R. (1991). Assessment literacy. *The Phi Delta Kappan, 72*(7), 534–539.

Stiggins, R. (1995). Assessment literacy for the 21st century. *The Phi Delta Kappan, 77* (3), 238–245. Retrieved from http://www.questia.com/read/1G1-17774629/assessment-literacy-for-the-21st-century

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21–36.

# Comparing Speaking Performances Across Tests and Languages: Evaluating the Success of an Institutional Rater-Training Program

**Koen Van Gorp**, Michigan State University, United States
**Daniel Reed**, Michigan State University, United States
**Susan Gass**, Michigan State University, United States
**Paula Winke**, Michigan State University, United States

**Abstract:** We evaluated the success of our institutional rater-training program to understand better if our scores are reliable and consistent for robust score interpretation and use. We analyzed the results of 59 college-level L2 language learners (19 Spanish, 19 French, and 21 Chinese) who took two separate speaking tests: an internal test and a national test rated by professional raters. We ran Pearson correlations between the 59 students' two sets of scores and compared the individual in-house raters' score assignments with those of the professional raters' ratings of the same students. We further assessed the consistency of the institutional raters' institutional test score assignments. We found our internal rater-training program needed improvement: The separate sets of test scores did not correlate highly, most likely due to inconsistent raters at the institutional level. We discuss how institutional rater-training programs can use such data analyses to evaluate and improve their operations.

## 1 Introduction

Speaking proficiency assessments play a central role in many language higher education programs across the United States. The Oral Proficiency Interview (OPI) of the American Council on the Teaching of Foreign Languages (ACTFL) is a widely- accepted standardized assessment. However, ACTFL speaking tests (face-to-face OPI; telephone-mediated OPIt; or computer-based OPIc) are not always the best solution for students and language programs, due primarily to practical issues. An in-house speaking assessment is often more practical and if done well can inform students and faculty about the students' oral proficiency and the extent to which program goals are being met. For an institution to have an in-house test that produces scores that are trusted as much as ACTFL tests, the institution must provide robust and iterative rater training and monitor the rating practices, as ACTFL does.

Participating in a 3-year U.S. government-funded initiative, Michigan State University (MSU), a large, public research institution, had the opportunity to compare its in-house ratings to official ACTFL ratings and, additionally, to test the rating quality of its in-house test. In this large-scale proficiency assessment project (July 2014 to June 2017), students studying Chinese, Russian, French and Spanish took ACTFL OPIcs, administered by Language Testing International (LTI), as part of their regular course requirements at the university. The purpose of this 3-year grant was to introduce proficiency assessment to established academic foreign language programs to measure teaching and learning, and to evaluate the impact of such testing (see Gass, Winke, & Van Gorp, 2016). A subset of students also took the Spanish, French or Chinese Simulated Oral Proficiency Interview (SOPI), which is used as an institutional proficiency test. Having both sets of data enabled us to evaluate the success of MSU's institutional test and rating program. We investigated how well the SOPI ratings aligned with the ACTFL ratings. We further investigated whether the SOPI rater program was producing scores as reliable/consistent as the ACTFL scores.

## 2 Context

Both OPIc and SOPI are rater-mediated oral proficiency assessments. Although the underlying speaking construct is operationalized in a slightly different manner, both tests aim to elicit a ratable sample of speech that represents the test takers' oral proficiency.

The ACTFL OPIc measures what language learners "can do with language in terms of speaking […] in real-world situations in a spontaneous and non-rehearsed context" (ACTFL, 2012, p. 3). This computer-based test imitates the face-to-face OPI through a computer program that lasts 20 to 40 minutes. Based on the test taker's self-assessment, which estimates Novice Low through Superior skills, one of five test forms is presented. Each form targets a different, but overlapping, proficiency range. For example, Test Form 4 tests proficiency levels Advanced Low through Advanced Mid, although a rating between Intermediate High and Advanced High can be assigned, depending on the amount of linguistic achievement and breakdown the speaker demonstrates. An avatar asks all questions. Two certified ACTFL raters rate the recorded speech by comparing the performances to the particular ACTFL (2012) proficiency levels targeted by the form. Overall, the ACTFL OPIc allows students to be tested at the Novice, Intermediate, Advanced, and Superior levels, with the first three levels subdivided into Low, Mid, and High.

Surface, Poncheri, and Bhavsar (2008) found the English OPIc to be a valid and reliable assessment of Korean students' oral proficiency: Its scores correlated highly with the OPI (Pearson's r = .92) in their study. The strength of this relationship is noteworthy because researchers comparing speaking sections of different academic English tests have generally found weaker relationships: Riazi (2013) compared the speaking section of the PTE Academic test to the *IELTS* test (r = .72); and ETS (2010) compared TOEFL iBT and *IELTS* speaking sections (r = .57). In addition, the OPIc allows for easy scheduling and administration.

The SOPI, developed by the Center for Applied Linguistics, is a performance-based speaking test that is similar to the OPIc. A newer computer-administered format has replaced the original cassette-tape format. Three parts (five picture tasks, five topics, and five situations) assess the examinee's ability to handle the functions and content characterizing ACTFL Intermediate, Advanced and Superior levels. The test lasts between 20 and 45 minutes. Two in-house raters rate the recorded speech.

Research on the SOPI, involving different contexts and target languages, has shown that the SOPI is a valid and reliable alternative to the OPI. Studies found Pearson correlations between .89 and .95: r = .93 for Chinese (Clark & Li, 1986), r = .89 for Hebrew (Shohamy, Gordon, Kenyon, & Stansfield, 1989), r = .93 for Portuguese (Stansfield, Kenyon, Paiva, Doyle, Ulsh, & Cowles, 1990), r = .95 for Indonesian (Stansfield & Kenyon, 1992), r = .91 for Hausa (Stansfield & Kenyon, 1993), and r = .94 for Spanish (Kenyon & Malabonga, 2001). To our knowledge, no study has compared the SOPI with the OPIc.

At MSU, language instructors are the ones who rate students' SOPI performances. For the last two academic years, raters received two training sessions per year. The first session lasted three hours and focused on the ACTFL Proficiency Guidelines, the SOPI tasks, and the rating of sample English performances. The second three-hour session focused on rating benchmarked target-language performances and reaching inter-rater consensus.

It is important for an institution of higher education that is responsible for evaluating its students' language requirements to demonstrate the reliability and validity of its SOPI's institutional use. Given that most students nowadays take the OPIc to meet their language requirement, evidence about how well the in-house SOPI ratings align with the external standard is needed.

## 3 Research questions (RQs)

RQ1: Are Spanish, French and Chinese students' in-house SOPI ratings equivalent to their ACTFL OPIc ratings?

RQ2: To what extent do individual SOPI raters vary in aligning their scores with OPIc scores?

RQ3: How consistently do the in-house SOPI raters score SOPI speech samples?

## 4 Methodology

The data include the test results from those students who took the SOPI and OPIc within a 4- to 6-week window. To investigate the rating quality, we ran correlations, inter-rater reliability, and rater agreement (Wind & Peterson, 2017).

### 4.1 Participants

Testing was conducted Spring 2015 and Spring 2016: 77 students studying Spanish (n = 26), French (n = 27) and Chinese (n = 24) at the third or fourth year of their undergraduate study took the SOPI; 59 of these (19 Spanish, 19 French, and 21 Chinese) also took the OPIc.

Five instructors of Spanish, three of French, and two of Chinese rated students' SOPI performances. One of the Spanish instructors, a prior French instructor, also rated the French performances. For the analyses, the raters received a number from 1 to 10 (R1 to R5 were raters of Spanish, R5 to R8 were raters of French, R9 and R10 were raters of Chinese). Four raters were experienced raters (R3, R4, R6 and R7); six were novice (R1, R2, R5, R8, R9 and R10). R4 and R5 were in the process of becoming ACTFL certified Spanish OPI raters. R2, R4, R9 and R10 were native speakers of the target language. All raters were (re)trained approximately one month before they actually rated the speech samples. Because we were interested in the consistency of the initial ratings of the raters involved, and not in the final agreement, no third ratings were applied.

### 4.2 Analyses

To explore students' SOPI and OPIc equivalencies, we computed Pearson product-moment correlation coefficients (r) between the SOPI and OPIc ratings for the three languages. Following Kenyon and Malabonga (2001), we first converted the proficiency ratings from Novice Low to Superior into a numerical score from 1 (Novice Low) to 10 (Superior). Second, we averaged and rounded down the two SOPI ratings to get one SOPI rating for each individual.

Additionally, we looked at Pearson's r between the SOPI and OPIc ratings for individual raters to see how consistent (well aligned) individual raters' SOPI ratings were with OPIc ratings.

To look at inter-rater reliability and rater variability, we computed Pearson correlations between each person's first and second rater and each rater's Intra-class Correlation (ICC). The ICC is a reliability measure that can be used when two or more raters are involved and rate different performances. It reflects the degree of correlation and measurement agreement (Koo & Mae, 2016). Larger magnitude disagreements result in lower ICCs than smaller magnitude disagreements. The ICC can be used to estimate the reliability of a single rater's rating. Koo and Mae (2016, p. 161) suggested that ICC values less than 0.5 indicate poor reliability, between 0.5 and 0.75 are moderate, values between 0.75 and 0.9 are good, and values over 0.90 are excellent. Besides the ICC estimate, its 95% confidence interval can reveal 'true' reliability (Koo & Mae, 2016).

## 5 Results

### 5.1 Correlations between SOPI and OPIc scores (RQ 1)

The Pearson correlation between the SOPI and OPIc scores for all examinees was r = .67 (df = 58, 95% CI [.49, .79], p < .001). The correlation between the Spanish SOPI and OPIc scores was r = .55 (df = 18, 95% CI [.13, .80], p = .015). The correlation between the French SOPI and OPIc scores was r = .53 (df = 18, 95% CI [.09, .79], p = .021). The correlation between the Chinese SOPI and OPIc scores was r = .80 (df = 20, 95% CI [.57, .92], p < .001).

### 5.2 Individual rater correlations between SOPI and OPIc scores (RQ2)

We found Pearson correlations between -.28 and .86 for raters of Spanish (R1: .86, R2: .73, R3: .26, R4: -.28 and R5: .83), between .14 and .92 for raters of French (R5: .92, R6: .82, R7: .48 and R8: .14), and a correlation of .82 between the two raters of Chinese (R9 and R10).

### 5.3 SOPI inter-rater reliability (RQ3)

The Pearson correlation between the first and second SOPI ratings for all examinees was .86 (df = 76, 95% CI [.79, .91], p < .001). The correlation was .76 (df = 25, 95% CI [.52, .88], p < .001) for the Spanish ratings, .85 (df = 26, 95% CI [.69, .93], p < .001) for the French ratings, and .94 (df = 23, 95% CI [.87, .97], p < .001) for the Chinese ratings.

The ICC (1,1) calculated on a one-way random effects model with 5 raters of Spanish across 26 subjects was .68 with a 95% confidence interval from .41 to .84 (F(25,26) = 5.305, p < .001). The same model with 4 raters of French across 27 subjects was .85 with a 95% confidence interval from .67 to .93 (F(26,27) = 12.122, p < .001). The ICC(2,1) calculated on a two-way random effects consistency model with 2 raters of Chinese over 24 subjects was .94 with a 95% confidence interval from .86 to .97 (F(23,23) = 31.809, p < .001).

## 6 Discussion

In this small-scale, exploratory study, we first looked at how well the ratings of an institutional oral proficiency test (SOPI) aligned with an external standard (ACTFL OPIc ratings). Second, we investigated individual raters' variability in assigning SOPI scores.

We found the Pearson correlations between SOPI and OPIc scores, except for the ratings of Chinese (r = .80), were rather low (r = .55 for Spanish; r = .53 for French). But they were not unlike the .57 correlation ETS (2010) found between the speaking sections of TOEFL iBT and *IELTS*. Nevertheless, these data seem to suggest that the SOPI ratings might not be robust student proficiency indicators at our institution, or, the OPIc and SOPI raters were not applying the same standards. We think the issue is the former because the analysis of how well the SOPI ratings of individual raters aligned with the OPIc ratings pointed to wide rater variability and uncovered inconsistent raters: R3 and R4 for Spanish and R7 and R8 for French. Interestingly, three of the four raters we identified as inconsistent have been rating SOPIs for many years (R3, R4 and R7). Only R8 was a novice rater. This adds evidence that these raters might be using criteria different from the criteria used by certified ACTFL raters (even though both tests are based on ACTFL Guidelines), and suggests that the SOPI rater training impacted experienced raters less.

Secondly, we investigated the inter-rater reliability of the SOPI ratings because rater consistency is key to high-quality ratings (American Education Research Association, American Psychological Association, & the National Council on Measurement in Education, 2014). Looking at the Pearson correlations, we conclude that our data showed acceptable (for Spanish), good (French), and excellent (Chinese) rater consistency. However, one limitation is that correlation ignores systematic differences in ratings. Therefore, we computed ICC to look at rater consistency as well as agreement. Using the ICC estimates and 95% confidence intervals, we uncovered large differences between SOPI raters, especially in Spanish, and, to a lesser extent, in French, confirming the finding of the correlations. Because the raters were not as consistent and reliable as expected, lower correlations between the SOPI and the OPIc scores were expected. The two go hand-in-hand.

Despite institutional efforts to raise rater-qualification requirements, and to improve rater training and norming sessions, rater inconsistency and misinterpretation of rating criteria often persist (e.g., Deygers & Van Gorp, 2015, Knoch, Read, & von Randow, 2007; Weigle, 1998). Wind and Peterson (2017) advised researchers to recognize these persistent problems and to incorporate information about individual raters into oral proficiency estimates. Rasch analysis or other scaled rating methods would allow for a better alignment of performance estimates and rater severity, and, therefore, could provide a more precise language proficiency measure. However, Rasch analysis might not be practical or even applicable for an in-house test with a low number of sporadic test takers and a required quick turnover of scores. Nevertheless, regular analyses of rater consistency and agreement will help researchers identify raters who agree with each other and align well with external standards. Giving well-performing raters a more prominent role in the in-house rating and rater-training program might be one step institutions can take toward improvement. A second step for us might be fine-tuning our training program not just to reach rater consensus, but also to align judgments with the ACTFL Proficiency Guidelines.

## 7 Conclusion

Analyzing the alignment of SOPI and ACTFL OPIc ratings provided valuable information on how our in-house testing related to the external ('gold') standard. Except for the Chinese ratings, our analyses pointed to some issues in our rating quality that need to be addressed. Test scores are indicators of students' language proficiency; the scores are interpreted and used for particular purposes. Consequently, rating quality is an inextricable part of test-score interpretation. Attention to the role of the rater in institutional performance assessment is needed, not just as a matter of reliability, but as a way to strengthen the validity argument of the in-house test and its use. The results push us (and we assume other institutions) to evaluate in-house training programs in more depth, in addition to conducting standard norming and monitoring practices, and to involve all stakeholders in the discussion about test score interpretation and use.

## References

ACTFL. (2012). *ACTFL Proficiency Guidelines 2012.* Retrieved from htttp://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/arabic

American Education Research Association, American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Clark, J.L.D. & Li, Y.C. (1986). *Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages*. Washington, DC: Center for Applied Linguistics.

Deygers, B. & Van Gorp, K. (2015). Determining the score validity of a co-constructed CEFR-based rating scale. *Language Testing, 32*(4), 521–541.

ETS (2010). *Linking TOEFL iBTTM scores to IELTS® scores—A research report*. Princeton: Educational Testing Service.

Gass, S., Winke, P., & Van Gorp, K. (2016). The Language Flagship Proficiency Initiative. *Language Teaching, 49(*4), 592–595.

Kenyon, D. M., & Malabonga, V. M. (2001). Comparing examinees' attitudes toward a computerized oral proficiency assessment. *Language Learning and Technology*, 5, 60–83.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26–43.

Koo, T.K. & Mae, Y.L. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Research, 15*, 155–163.

Riazi, M. (2013). Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic). *Papers in Language Testing and Assessment, 2*(2), 1–27.

Shohamy, E., Gordon, C., Kenyon, D.M., & Stansfield, C.W. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Hebrew Higher Education, 4*, 4–9.

Stansfield, C.W., & Kenyon, D.M. (1992). The development and validation of a simulated oral proficiency interview. *Modern Language Journal, 76*, 129–141.

Stansfield, C.W., & Kenyon, D.M. (1993). Development and validation of the Hausa Speaking Test with the ACTFL Proficiency Guidelines. *Issues in Applied Linguistics, 4*, 5–31.

Stansfield, C.W., Kenyon, D.M., Paiva, D., Doyle, F., Ulsh, I., & Cowles, M.A. (1990). Development and validation of the Portuguese Speaking Test. *Hispania, 73*, 641–651.

Surface, E., Poncheri, R., & Bhavsar, K. (2008). *Two studies investigating the reliability and validity of the English ACTFL OPIc with Korean test takers: The ACTFL OPIc validation project technical report*. Retrieved from http://www.languagetesting.com/wp-content/uploads/2013/08/ACTFL-OPIc-English-Validation-2008.pdf.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287.

Wind, S.A. & Peterson, M.E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, 1–32.

# Exploring Teachers’ Language Assessment Literacy: A Social Constructivist Approach to Understanding Effective Practices

**Vivien Berry**, British Council, UK
**Susan Sheehan**, University of Huddersfield, UK
**Sonia Munro**, University of Huddersfield, UK

**Abstract:** Exploring teachers’ levels of assessment literacy in terms of their previous assessment experiences may help teacher educators to better understand the factors which promote or prevent effective assessment, thus contributing to more targeted and empowering teacher education.

The research presented in this paper adopts a social constructivist model of learning and meaning-making, with the language classroom representing the community of practice. The first phase of the project consisted of interviews with teachers, in which they were invited to estimate their understanding of individual components of the assessment process and indicate how much they would like to learn about each. Classroom observations then took place followed by post-observation, reflective interviews. Finally, focus group discussions were conducted with further groups of experienced teachers.

Four key findings are presented, highlighting the considerable differences in understanding which exist between teachers and those who research and write about language testing/assessment.

## 1 Introduction

Jones & Saville (2016) assert that the two key purposes of assessment are to promote learning and to measure and interpret what has been learned. In terms of classroom assessment, this implies that teachers have a central role to play in planning and/or implementing appropriate assessment procedures to monitor and evaluate student progress in their classrooms. But teachers’ attitudes and beliefs, based on their own experiences of assessment, exert a powerful role in shaping their decisions, judgements and behaviour (Borg, 2006; Kagan, 1992). Consequently, exploring teachers’ levels of assessment literacy in terms of their own assessment experiences may help teacher educators to better understand the factors which promote or prevent effective assessment, thus contributing to more targeted and empowering teacher education.

Assessment literacy has been defined by numerous researchers in broadly similar ways. According to O’Loughlin (2013), language assessment literacy “potentially includes the acquisition of a range of skills related to test production, test score interpretation and use, and test evaluation in conjunction with the development of a critical understanding about the roles and functions of assessment within education and society” (p. 363). Pill & Harding (2013) offer a succinct overview “language assessment literacy may be understood as indicating a repertoire of competences that enable an individual to understand, evaluate and, in some cases, create language tests and analyse test data” (p. 381). Fulcher’s definition is more complex and draws to a large extent on Davies’ (2008) components of assessment literacy, skills, knowledge and principles. For Fulcher (2012), assessment literacy consists of:

> The knowledge, skills and abilities required to design, develop, maintain or evaluate large-scale standardized and/or classroom based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals. (p. 125).

Malone (2011) offers a definition which is specifically relevant to a focus on classroom teachers' language assessment literacy: "Assessment literacy is an understanding of the measurement basics related directly to classroom learning; language assessment literacy extends this definition to issues specific to language classrooms."

## 2 Previous research into language assessment literacy

### 2.1 Survey research

Much previous research into teachers' assessment literacy has relied on survey data (Berry and O'Sullivan, 2014; Brown and Bailey, 2008; Crusan, Plakans and Gebril, 2016; Fulcher, 2012; Hasselgreen, Carlsen and Helness, 2004; Jin, 2010; Kiomrs, Abdolmehdi and Naser; Malone, 2013, inter alia). However, although they provide a valuable tool for collecting large amounts of data quickly, and from a wide geographical constituency if delivered and completed online, survey studies have several limitations.

First, respondents to surveys, especially online surveys, are probably self-selected as those who are interested in the topic in the first place. Second, teachers' responses may reflect what they think they should say, rather than what they actually believe. A corollary to this is that training needs may be exaggerated in the belief that it would appear unprofessional to state that they had no interest in the topic. Also, affirmative answers may be given out of curiosity rather than genuine interest or need to know. And following data collection, interpretation of responses may rely too heavily on quantitative analysis at the expense of individual differences.

### 2.2 Mixed methods and classroom observation studies

Stoynoff (2012) claims that "survey results need to be complemented with other empirical evidence of the effect of teacher characteristics on assessment practices" (p. 531). Several studies have attempted to address this statement either through the use of mixed methods or through classroom observation (Colby-Kelly and Turner, 2007; Gu, 2014; Jeong, 2013; Lam, 2015; Leong, 2014; Scarino, 2013; Vogt and Tsagari, 2014; Xu and Carless, 2016; Xu and Liu, 2009; Yin, 2010). While attempting to take a more qualitative approach to data collection than survey studies, mixed methods and classroom observation studies also have limitations.

Mixed methods studies generally include initial questionnaire/survey responses as a basis for follow-up interviews. These follow-up interviews usually aim to elicit further insights into responses given to the questionnaires/surveys and therefore all limitations that apply to survey studies are also applicable to mixed methods studies. Responses are also likely to be constrained by the questions asked by the interviewer. Qualitative aspects of mixed methods and, in particular, classroom observation studies are likely to be very small. For example, mixed methods studies usually only ask a small percentage of respondents to participate in follow-up interviews and in the case of Xu and Liu (2009), who used narrative enquiry to explore teacher assessment knowledge and practice, their study was of one person. These small-scale studies make it impossible to generalise the findings beyond the immediate participant population.

## 3 Methodology

### 3.1 Aims of the study

Having identified several limitations concerning both survey and mixed methods studies, this study adopts a social constructivist model of learning and meaning-making, with the language classroom representing the community of practice. It focuses on the sociocultural context in relation to actual assessment literacy practices in the language classroom, since an investigation into what is happening in classes may be of little value without exploring why it is happening. With the exception of the case study mentioned above, which followed one Chinese University teacher (Xu and Liu, 2009), no teachers have been asked specifically about their attitudes to assessment or their actual training needs. This study aims to bring teachers more directly into the assessment literacy debate in order to provide them with training materials which meet their actual stated needs.

### 3.2 Participants in the study

The study consisted of 3 phases. In Phase 1, three experienced international EFL teachers, 2 male + 1 female, age range 30–50 years, were interviewed. In Phase 2, three teachers, 2 female and 1 male, age range 30–40 years, were observed in their classrooms. The observed teachers were not the same as the ones interviewed in Phase 1. In Phase 3, 48 experienced EFL teachers, 25 female, 23 male, age range 25-60 years, participated in 5 focus group discussions.

### 3.3 Procedures

#### *3.3.1 Phase 1 – interviews*

The first phase of the project consisted of a series of interviews with three experienced, international EFL teachers, conducted in the School of Education of a British university. The interviews drew on Davies' (2008) components of assessment literacy which he defined as skills + knowledge but with the important addition of principles. These components can be summarised as technical skills, scores and decision making, language pedagogy, local practices, knowledge of theory, principles and concepts, socio-cultural values, personal beliefs and attitudes.

Teachers were asked about their experiences of assessment and how they had developed their assessment practices. They also discussed their initial teacher training and other training opportunities they had had. In the interviews, teachers were invited to estimate their understanding of the components of the assessment process and indicate how much they would like to learn about each individual component.

#### *3.3.2 Phase 2 – observations and follow-up interviews*

In the second phase, observations were conducted in the International Study Centre of a British university which focused on teachers' actual assessment practices in the classroom. Using an observation schedule inspired by Colby-Kelly and Turner's (2007) study of assessment for learning practices, we developed a checklist of 16 assessment practices and every 3 minutes during the observations, checked which of the practices were being observed and took notes about them. Post-observation interviews were subsequently conducted with the 3 teachers, in which they were asked to reflect on their observed classroom practice and discuss why they had used particular assessment techniques in class.

#### *3.3.3 Phase 3 – focus group discussions*

Finally, focus group discussions were held with 48 experienced teachers working at teaching centres attached to a major international organisation in Madrid and Paris. These teachers taught a variety of different English language classes across a range of ages and proficiency, including kindergarten, elementary, secondary and tertiary level students, plus special-purpose classes for commercial organisations. The group interview schedule also drew on Davies' (2008) components. These discussions confirmed the findings from the initial phase of the project, culminating in the creation of a set of online training materials.

## 4 Findings

The data analysis drew on Davies' (2008) components of assessment literacy as detailed in section 3.3. Three key findings emerged from the analysis relating to teachers' previous training in assessment, attitudes to language testing and assessment in its broader sense and the types of training materials they would like. Regarding the quotations from the teachers below, those who participated in the baseline interviews are referred to as IT, those who were observed and interviewed are referred to as OT and the focus group participants are referred to as FGT.

### 4.1 Previous training in assessment

Davies' (2008) components, skills + knowledge + principles, was only used as a data code on 12 occasions. In discussion, teachers acknowledged their lack of training, exemplified in the following quotations:

> FGT9: There are so many things that I didn't have a clue about how to do so I wouldn't put assessment at the top of my list
> OT1: We were not planning and designing assessments we were planning and delivering lessons
> IT2: We didn't do it (assessment) in practice on the assessments

It may be the case that the divide between teaching and assessment starts to develop at pre-service training. Teaching is prioritised and assessment is not considered to be important.

### 4.2 Attitudes to language testing and assessment

In discussion, participants tended to refer to testing rather than assessment. In our questions the word assessment was used. The following are representative quotations:

> IT1: None of my experiences of teaching had any focus on any kind of qualification at the end of it
> FGT20: The idea of grading someone isn't that important
> FGT35: You need to understand the exam techniques to prepare students to take exams
> OT2: In most places testing and assessment is out of the hands of teachers …. they are told 'this is the assessment you are using'
> FGT13: Assessment requires some level of experience with students
> IT3: If I have read any books about language testing it was from the perspective of being interested in researching the language classroom and sometimes in classroom research you need tests

FGT4: You build up your own ideas of assessment just through experience of what your students are capable of doing

FGT24: You bring conceptions of how you were tested at school and you apply them to the language classroom

The lack of engagement with assessment may be a consequence of the limited role some teachers play in the development and creation of assessments. This would seem to provide support for the notion that teachers feel assessment is a top-down imposition (Crusan et al., 2016). In addition, there is some evidence to suggest that testing is only acceptable if it can be used to support or improve teaching in some way. This is a further demonstration of the gap between teaching and assessment as teaching is being privileged. Experience, rather than training, seems to play a pivotal role in the development of assessment practices. This experience develops with time spent in the classroom. There also seems to be evidence to suggest that experiences in the classroom as school children influence how teachers develop practices relating to assessment. This brings notions of the “apprenticeship of observation” (Lortie, 2002) to the fore. All trainee teachers have experienced thousands of hours in the classroom before they start teaching. It is, perhaps, not surprising, that practices experienced as a school child contribute to the formation of assessment practices in teachers as adults. This idea is deemed to be problematic by Vogt and Tsagari (2014). They make the analogy between “teaching as you were taught” and “testing as you were tested”. This is characterised as a brake on innovation and a hindrance to the development of effective assessment practices.

### 4.3 Types of training materials requested

Most of the teachers who participated in the study expressed their training needs in terms of requests for activities and not in terms of theory or principles, thus confirming Davies’ (2008) claim that there is little demand for theory among teachers. Teachers mainly requested training materials related to skills, replicating, in the main, the findings of Berry and O’Sullivan (2014) and Hasselgreen et al. (2004), as exemplified in the following quotations:

FGT7: We’d like speaking tasks – tasks and criteria

FGT2: We’d like clear criteria for marking speaking and writing

FGT45: Examples of level – recordings or writings from non-exam classes

FGT24: Video examples of people in everyday situations using the language

IT2: I would have liked more practical elements in my training and assessment – more situation based

These quotations suggest that the training they had received did not prepare the teachers for the type of assessments they engage in. They may also be an indication of how busy the teachers are and that they lack time to develop assessments. Coombe, Troudi and Al-Hamly (2012) suggest that teachers avoid engaging with assessment as they do not have access to adequate assessment resources.

### 4.4 Overall finding

The term Language Assessment Literacy was not popular with teachers and many were not even familiar with it.

FGT40: I had never heard of it before I was asked to do the interview

FGT5: I have no idea what it means

The term has been widely used in the language assessment literature but, on the evidence of this project, has not entered into teacher language.

## 5 Conclusions

It would seem from the interviews, observations and focus group discussions that teachers have minimal training in assessment and have little interest in the theoretical underpinnings of assessment. There is evidence that teachers' assessment practices are rooted in their own past learning experiences, confirming the claims of Borg (2006) and Kagan (1992). Teachers also develop their assessment practices over time by learning from each other.

It may also be that there is a disconnect between teachers' interests and beliefs and those of language assessment professionals and researchers. Our findings suggest that the gap between teachers and those who research and write about language testing is considerable. This project sought to narrow the gap by giving teachers a stronger voice in the debate, which, in turn, may have important implications for the development of future teacher training courses.

## References

Berry, V. & O'Sullivan, B. (2014). *The symbiosis of teachers' language assessment literacy and learning-oriented outcomes*. Paper presented at IATEFL TEASIG Conference, October 2014, Granada, Spain

Borg, S. (2006). *Teacher cognition and language education: Research and practice*. London: Continuum.

Brown, J. D. & Bailey K. M. (2008). Language testing courses: What are they in 2007? *Language Testing, 25*(3), 349–384.

Colby-Kelly, C. & Turner, C. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *The Canadian Modern Language Review, 64*(1), 9–37

Coombe, C., Troudi, S., & Al-Hamly, M. (2012). Foreign and second language teacher assessment literacy: Issues, challenges and recommendations. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 20–29). Cambridge: Cambridge University Press.

Crusan, D, Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28*, 43–56.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327–347.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

Gu, P. (2014). The unbearable lightness of the curriculum: What drives the assessment practices of a teacher of English as a foreign language in a Chinese secondary school. *Assessment in Education: Principles, Policy and Practice,21*(3), 286–305.

Hasselgreen, A., Carlsen, C., & Helness, H. (2004). *European survey of language testing and assessment needs: General findings*. Retrieved from: www.ealta.eu.org/resources.htm

Jeong, H. (2013). Defining assessment literacy: Is it different for language testers and non-language testers? *Language Testing, 30*(3), 345–362.

Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing, 27*(4), 555–584.

Jones, N. & Saville, N. (2016). *Learning Oriented Assessment: a systemic approach*. Studies in Language Testing 45. Cambridge: UCLES/Cambridge University Press.

Kagan, D. M. (1992). Implications of research on teacher belief. *Educational Psychologist, 27*(1), 65–90.

Kiomrs, R., Abdolmehdi, R., & Naser, R. (2011). On the interaction of test washback and teacher assessment literacy: the case of Iranian EFL secondary school teachers. *English Language Teaching, 4*(1), 156–161.

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing, 32*(2), 169–197.

Leong, W. S. (2014). Knowing the intentions, meaning and context of classroom assessment: A case study of Singaporean teachers' conception and practice. *Studies in Educational Evaluation, 43*, 70–78.

Lortie, D. C. (2002). *Schoolteacher: A sociological study*. Chicago: University of Chicago Press.

Malone, M.E. (2011). *Assessment Literacy for Language Educators. CAL Digest October 2011*. Retrieved from http://www.cal.org

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344.

O'Loughlin, K. (2013) Developing the assessment literacy of university proficiency test users. *Language Testing 30*(3), 363–380.

Pill, J. & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary Inquiry. *Language Testing, 30*(3): 381–402.

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309–327.

Stiggins, R. (2014). Improve assessment literacy outside of schools too: Teaching and assessment have become separated, which has kept teachers from developing the assessment skills they need to truly enhance learning. *Phi Delta Kappan, 96*(2), 67–72.

Stoynoff, S. (2012) Looking backward and forward at classroom-based assessment. *ELT Journal, 66*(4), 523–534.

Xu, Y. & Carless, D. ((2016) Only true friends could be cruelly honest: Cognitive scaffolding and social-affective support in teacher feedback literacy. *Assessment & Evaluation in Higher Education*, 1–13.

Xu, Y. & Liu, Y. (2009) Teacher assessment knowledge and practice: A narrative inquiry of a Chinese college EFL teacher's experience. *TESOL Quarterly,43*(3), 492–513.

Vogt, K. & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402.

Yin, M. (2010). Understanding classroom language assessment through teacher thinking research. *Language Assessment Quarterly, 7*(2), 175–194.

# Construction of Attribute-based Dynamic Mediation for Cognitive Diagnostic EFL Listening Test

**Yihe Yan**, Xi'an Jiaotong University, China
**Xiaomei Ma**, Xi'an Jiaotong University, China

**Abstract:** Though at present there is a growing support for the use of DA in second language pedagogy, few studies focus on L2 listening and most researches are small-scale case studies. The present study aims to construct attribute-based mediation and specifically presecribed prompts based on the cognitive diagnostic EFL listening test, hoping to provide a framework of mediation that can be used in Computerized-DA contexts. Both qualitative and quantitative methods are employed. The attributes to be mediated are first determined by R software, and designed through literature reviewing, meta-analysis and students' think-aloud protocols (TAPs). Through semi-structured interviews, most students confirm the usefulness of the mediation, and think the mediation could help them better comprehend the listening text and direct their attention to their listening deficiencies.

## 1 Introduction

The relationship between instruction and assessment is a longstanding concern in applied linguistics (Poehner, 2005). Traditionally, teaching and assessment remain dichotomised with the former focusing on learning and the latter measuring on that learning. The purpose of assessment primarily lies in assessing learners' abilities rather than promoting learner development. Getting high scores is considered so important that it leads to some unfavourable consequences like "teaching to the test", and assessments actually stand in the way of instructional practices. To change this situation, researchers have striven to reach a more fine-grained dignosis so as to truly inform teaching and learning. Cognitive Diagnostic Assessment (CDA), by measuring specific knowledge structures and processing skills to provide formative diagnostic feedback about learners' cognitive strengths and weaknesses (Leighton & Gierl, 2007), can better achieve the diagnostic purposes. Nevertheless, it is a kind of static and non-interactive diagnostic procedure in which teaching and assessment still remain as separated activities.

Unlike traditional view that instructions are not permitted in assessment because they would obscure the learners' true abilities, Dynamic Assessment (DA) argues that instruction and assessment must be unified into a single activity in which various forms of support are provided to reveal the scope of learners' abilities while simultaneously aiding their development (Lidz & Gindis, 2003).

In recent years, DA has been increasingly applied in L2 studies. And Computerized-Dynamic Assessment (C-DA) is gaining more and more popularity (Bjrjandi & Ebadi, 2011; Ebadi, 2016; Khonamri & Sana'Ati, 2014; Poehner & Lantolf, 2013; Wang, 2010). Researchers always applied an interventionist approach in C-DA since it can be simultaneously administered to large numbers of learners and is more time and energy-saving. However, most L2 interventionist DA studies focus on speaking, reading and writing skills (Davin, 2013; Kozulin & Garb, 2002; Poehner & Lantolf, 2010), with few studies focusing on learners' listening comprehension. What's more, though many attempts have been made, there are still some areas that need to be improved. Firstly, the validity and reliability of the diagnostic tool used haven't been thoroughly justified, so the dependability of the testing results is challenged to some extent. Secondly, the success of Computerized-DA hinges upon preparing mediation that is as responsive to learner needs as is feasible in the absence of co-regulation (Lantolf & Poehner, 2014). However, the mediation offered in C-DA fails to fully target the construct and the online format also limits its form and content, making it difficult to meet learners' needs. Therefore, the present study aims to maximise the positive influence of C-DA, both in the test, and the mediation offered.

## 2 Theoretical framework of dynamic assessment

Vygotsky's Social Cultural Theory (SCT) provides theoretical and methodological underpinnings for the present study. Grounded in Vygotsky's Zone of Proximal Development (ZPD), DA has evolved into an assessment with varied definitions and formats. The most frequently referenced definition of the ZPD is "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978). Mediation in itself is a simple concept but has tremendous consequences for individuals' intellectual development, which has its beginning in the ZPD (Ableeva & Lantolf, 2011). Mediation is considered as the major means to realize DA. The SCT framework understands mediation as: "the process through which humans deploy culturally constructed artifacts, concepts, and activities to regulate (i.e. gain voluntary control over and transform) the material world or their own and each other' s social and mental activity" (Lantolf & Thorne, 2006), and defines DA as follows: it integrates assessment and instruction into a seamless and unified activity aimed at promoting learner development through appropriate forms of mediation that are sensitive to the individuals' (or in some cases a group's) current abilities. In essence, DA is a procedure for simultaneously assessing and promoting development that takes account of the individual's (or group's) ZPD (Lantolf & Poehner, 2004).

The Mediated Learning Experience (MLE), developed by Reuven Feuerstein and his colleagues, advocates open-ended dialogue to reveal underlying difficulties and to begin the process of mediating development (Poehner, Zhang, & Lu, 2015). During MLE, an adult mediator engages in a task with a learner and provides as much mediation and as many forms of mediation as necessary to improve performance. One aim of MLE is to determine the forms of mediation to which individual learners appear most responsive (Sternberg & Grigorenko, 2002) and elicited verbalizations from learners also provide insights into the nature of the problems they experienced. Thus in the present study, MLE is used to help the author design and modify the attribute-based mediation.

The Graduated Prompt Approach (GPA) developed by Campione and Brown is a variation on interventionist DA. A hierarchy of graduated prompts is offered, arranged from most implicit to most explicit. During the administration of the test, whenever a learner encounters difficulties, the mediator begins with the most implicit prompt and moves step by step toward the explicit ones until the learner correctly answer the question or until the mediator finally reveals the solution and explains why it is correct (Ableeva, 2010). In the present study, the designed mediation will be used in the final stage by applying GPA.

## 3 Research purpose and questions

The web system, Personalized English Learning: Diagnosis and Guidance (PELDiaG), was developed by the research team from Xi'an Jiaotong University based on the theories of Cognitive Diagnostic Assessment (CDA). Within this system, a cognitive diagnostic EFL listening comprehension model has been constructed, with three cognitive diagnostic listening tests and its corresponding item/attribute Q-matrices included. Seven listening attributes have also been identified, which are phonological features, vocabulary and expressions, special structures, facts and details, main ideas, and context- and culture-based inference.

This study aims to construct attribute-based mediation for one of the cognitive diagnostic listening tests in PELDiaG. Its ultimate purpose is to provide a framework of attribute-based mediation for an online CDA-based Dynamic Intervention system.

This study intends to address the following three research questions (RQs):

1) How can the attributes used for mediation be identified?

2) How is attribute-based mediation constructed?

3) To what extent is the attribute-based mediation reliable and valid?

## 3 Research design and methodology

### 3.1 Research design

In order to construct attribute-based mediation, this study applied both qualitative and quantitative approaches to conduct the experiment and to report its results.

The research design is as follows: large-scale data collection through non-dynamic assessment 1 (NDA1), think-aloud protocols (NDA2), case studies (DA1), an interventionist approach to DA (DA2) and semi-structured interviews. For NDA1, quantitative and statistical analysis was used to determine the final version of the diagnostic test and the attributes to be mediated. For NDA2 and DA1, a qualitative approach was applied to interpret the data, helping the author to design, modify and validate the attribute-based mediation. Then DA2 and interviews served to investigate learners' perceptions of the designed mediation. Each session had its corresponding examinees, and participants were recruited separately.

### 3.2 Instrument

Different test items were chosen from the aforementioned three diagnostic tests in the PELDiaG system, and some adaptations were made to refine the test items. As a result, the listening test adopted was a five-option format for multiple-choice questions. It is composed of three sections, containing 22 items, with specific cognitive attributes examined by each item.

### 3.3 Participants

The participants involved in the current study were divided into four groups. All the participants were Chinese native speakers, ranging in age from 18 to 25. First, a large-scale diagnostic listening test was administered separately within six different universities in China, involving 1,121 students in total. Then 12 students were invited for think-aloud protocols and five for case studies, both providing a basis for the construction of mediation. Finally, 12 students were involved in the interventionist DA session, and were all interviewed to investigate their perceptions towards the designed mediation.

### 3.4 Procedure

In the preparation stage, test items were selected and modified by experts and peer researchers, then coded by the eight experts. A large-scale test was administered in six different universities. The coding results along with students' response data were analyzed to help the author determine the final version of the diagnostic test, and the attributes to be mediated for each item were identified by using R software.

In the second stage, first, a hypothetical framework of mediational moves ranged from most implicit to most explicit was designed based on relevant literature reviewing. Then TAP and case studies were carried out to help the author modify and validate the attribute-based mediation, which was actually an iterative design and modification process. For all experiments, the test was presented in a paper format, and examinees were required to listen to the radio and then to choose the best answer from the five choices. For case studies, the previously designed moves were used to help examinees when they encountered diffculties, but the moves were not followed strictly; instead, they were applied flexibly. Experiments were all audio-recorded, transcribed and coded. Transcripts were classified into the specific mediational moves and the most frequently used. As a result, the effective moves were identified, which were incorporated in the interventionist DA session. After the large-scale interventionist DA session, semi-structured interviews were conducted to investigate examinees' perceptions of the mediation.

## 4 Discussion and conclusion

Dynamic Assessment has been introduced in China for more than a decade, but most researches are confined in classroom and experimental-developmental contexts, which are of low efficiency and quite time-consuming. Up till now, few large-scale researches have been carried out. The prime concern of this study is to design prescribed attribute-based mediation, which can be used in large-scale Computerized-DA. After the iterative and modification process, the designed mediation was used in interventionist DA. Most examinees found the attribute-based mediation quite useful in helpig them comprehend the listening text. Some reflected that the mediation could help them shift their attention to the examined attributes or construct, narrowing their attention span. Some thought the text is more useful than the clipped video, while some thought vice versa. No matter which part of the mediation contributes to their understanding of the listening text, it is well-received by most students. However, there are still some problems need to be tackled. The identified attributes may not be the only reason that led to incorrect answer. Different students may encounter different problems when answering the same questions, what’s more, students with different ability levels also show different problems when answering questions. Thus when designing mediation, we’re required to take different students’ needs into consideration so as to better promote their listening abilities.

The significance of the current study lies in two aspects. Theoretically, this study focuses on both EFL learners’ knowledge structures and processing skills emphasized by cognitive psychologists, and the dynamic intervention emphasized by social cultural theory. By applying both diagnostic tests and procedures, we aim at truly forming an organic whole of diagnosis, feedback, and intervention. Practically, as a preliminary attempt to design the attribute-based mediation, it can open up a new perspective for researchers to design mediating prompts. Meanwhile it can reach a fine-grained diagnosis and inform subsequent teaching and learning. The ultimate purpose of this study is to provide a framework for the construction of an online CDA-based Dynamic Intervention system.

## References

Ableeva, R. (2010). Dynamic Assessment of Listening Comprehension In Second Language Learning. *Procedia – Social and Behavioral Science, 98*, 1,729–1,737.

Ableeva, R. & Lantolf, J. (2011). Mediated dialogue and the microgenesis of second language listening comprehension. *Assessment in Education: Principles, Policy & Practice*, *18*(2), 133–149.

Birjandi, P. & Ebadi, S. (2012). Microgenesis in dynamic assessment of l2 learners’ socio-cognitive development via web 2.0. *Procedia - Social and Behavioral Sciences, 32*, 34–39.

Davin, K. J. (2013). Integration of dynamic assessment and instructional conversations to promote development and improve assessment in the language classroom. *Language Teaching Research, 17*(3), 303–322.

Ebadi, S. (2016). Exploring DIALANG's diagnostic feedback in online L2 dynamic assessment. *Teaching English with Technology, 16*(1),41–58.

Khonamri, F. & Sana’Ati, M. K. (2014). The impacts of dynamic assessment and call on critical reading: an interventionist approach. *Procedia - Social and Behavioral Sciences, 98*, 982–989.

Kozulin, A. & Garb, E. (2002). Dynamic assessment of efl text comprehension. *School Psychology International, 23*(1), 112–127.

Lantolf, J. P. & Poehner, M. E. (2004). Dynamic assessment of L2 development: bringing the past into the future. *Journal of Applied Linguistics,1*(2), 49–72.

Lantolf, J. P. & Poehner, M. E. (2014). *Sociocultural theory and the pedagogical imperative in L2 education: Vygotskian praxis and the theory/research divide*. London: Routledge.

Lantolf, J.P. & Thorne, S.L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford: Oxford University Press.

Larsen-Freeman, D. & Cameron, L. (2008). *Complex Systems and Applied Linguistics*.Oxford: Oxford University Press.

Leighton J.P., Gierl M.J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees’ thinking processes.*Educational Measurement: Issues and Practices, 26*(2), 3–16.

Lidz, C. S. & Gindis, B. (2003). Dynamic assessment of the evolving cognitive functions in children. In B. Kozulin, V. S. Gindis, S. Ageyev, & M. Miller (Eds.),*Vygotsky’s educational theory in cultural context* (99–116).

Cambridge: Cambridge University Press.

Kozulin, B., Gindis, V. S., Ageyev, S., & Miller, M. (Eds.). *Vygotsky’s educational theory in cultural context*.

Cambridge: Cambridge University Press.

Poehner, M. E. (2005). *Dynamic assessment of oral proficiency among advanced L2 learners of French*. Unpublished doctoral dissertation. The Pennsylvania State University, University Park, PA.

Poehner, M. E., Zhang, J., & Lu, X. (2015). Computerized dynamic assessment (CDA):Diagnosing L2 development according to learner responsiveness to mediation. *Language Testing, 32*(3) 337–357.

Poehner, M. E & Lantolf, J. P. (2010). Vygotsky's teaching-assessment dialectic and L2 education: The case for dynamic assessment*. Mind, Culture, and Activity*,17, 312–330.

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: capturing L2 development during computerized dynamic assessment (CDA). *Language Teaching Research, 17*(3), 323–342.

Sternberg, R. J. & Grigorenko, E. L. (2002). *Dynamic testing. The nature and measurement of learning potential*. Cambridge: Cambridge University Press.

Vygotsky, L. S. (1978). *Mind in society. The development of higher psychological processes*. Cambridge: Harvard University Press.

Wang, T. H. (2010). Web-based dynamic assessment: taking assessment as teaching and learning strategy for improving students' e-learning effectiveness. *Computers & Education, 54*(4), 1,157–1,166.

# Integrating Corpus Linguistics and Classroom-based Assessment: Evidence from Young Learners’ Written Corpora

**Trisevgeni Liontou**, Ministry of Education, Research and Religious Affairs, Greece
**Dina Tsagari**, University of Cyprus, Cyprus

**Abstract:** This study reports on a 1-year longitudinal study that assessed EFL young learners’ writing proficiency as a function of grade level. A total of 50 EFL students aged 9 to 11 years old produced 500 written essay that were analysed through a range of advanced computational linguistics and automated machine learning systems. The analysis was based on a number of linguistic features related to propositional density, lexical sophistication, syntactic complexity, and cohesion. The results showed statistically significant differences between the linguistic features at different levels of language competence of the young EFL learners. The findings support that linguistic development occurs in the later stages of writing development and is related to more elaborate texts with more sophisticated words, more complex sentence structure, and fewer cohesive features as a function of grade level. The study concludes by providing practical guidance to EFL teachers, curricula and assessment designers.

## 1 Introduction

Over the past 30 years, analysing writing development across grade levels has provided writing researchers with crucial information about how writing skills change as neural, cognitive, and linguistic functions develop (Ferris, 2004; Crossley & McNamara, 2009; Crossley, Salsbury, McNamara, & Jarvis, 2010; Shaaban, 2000; Silva, 1993; Uhl-Chamot & El-Dinary, 1999). The results showed that analysing writing development as a function of grade level is important in elementary school children because the developmental patterns are stronger at a young age when successful interventions are needed (Berninger, Cartwright, Yates, Swanson, & Abbot, 1994; Haswell, 2000; McNamara, Max, McCarthy, & Graesser, 2010; Witte and Faigley, 1981).

Motivated by the above literature on writing development, a one-year longitudinal study was carried out in order to assess EFL young learners’ writing proficiency as a function of their grade level. In order to explore this issue, the following research questions were formed: Are there any statistically significant lexicogrammatical differences between essays written by young EFL learners at different grade levels? If yes, which text variables can better predict text complexity variation between grade level students?

## 2 Methodology of the study

### 2.1 Participants and written corpus

Students (N = 50) came from a primary school located in Athens, Greece. Half of the participants were 4th grade students (9 years old) and the remainder were 5th grade students (11 years old). At the time of the study they had all been learning English as a Foreign Language as a compulsory school subject for a minimum of 3 years. Their language proficiency was diagnosed through a calibrated English language test (*Cambridge English: Key* – KET). Throughout the school year, students were kindly requested to produce a variety of written texts based on a specific set of descriptive prompts included in their school coursebook (e.g. describe your school, hobbies, neighbourhood, etc). A total of 500 written essays were collected, with a fixed number of 10 written essays per student.

### 2.2 Linguistic text analysis

135 text variables were chosen for both practical and theoretical reasons, e.g. the presence of cohesive ties created by referencing and conjunction and lexical cohesion; nominal group structure; grammatical intricacy; lexical density; surface text features (i.e., the number of words, sentences, and paragraphs per text, word frequency, lexical diversity, propositional density, proportion of passive sentences, negations, phrasal verbs, and idioms per text), etc (see Liontou and Tsagari, 2016). Also, estimates from 4 well-known readability formulas (i.e., the Flesh Reading Ease Index, the Dale-Chall Readability Index, the Fry Readability Index, and the Gunning-Fog Index) was determined. Finally, IBM SPSS 20.0 statistical package data was used to compute descriptive statistics and perform Pearson product moment correlations and T-tests.

### 2.3 Automated text analysis tools

In the present study, Coh-Metrix 2.1, Linguistic Inquiry and Word Count 2007 (LIWC), the VocabProfile 3.0, Computerized Language Analysis (CLAN) suite of programs, Computerized Propositional Idea Density Rater 3.0 (CPIDR), Gramulator, and TextAnalyzer were used to estimate the text variables.

## 3 Findings and discussion

Once the analysis of text characteristics per grade level was completed, independent sample t-tests were carried out in order to explore and further determine the significance of existing differences between 4th grade and 5th grade written texts.

| | **4th Grade** | **5th Grade** | | | |
|---|---|---|---|---|---|
| | **Mean** | **Mean** | **t** | **df** | **Adj. sig.** |
| 1. Words/Sentence | 9.04 | 10.10 | -2.357 | 498 | .019 |
| 2. Syllables/Word | 1.32 | 1.35 | -2.153 | 498 | .032 |
| 3. Letters/Word | 3.95 | 4.02 | -1.996 | 498 | .042 |
| 4. Syntactic simplicity | 44.98 | 55.01 | 3.900 | 498 | .001 |
| 5. Word concreteness | 74.60 | 71.00 | 1.339 | 498 | .006 |
| 6. Referential cohesion | 80.46 | 68.91 | 5.153 | 498 | .001 |
| 7. Connectivity | 5.94 | 10.01 | -2.397 | 498 | .001 |
| 8. Lexical diversity | 36.71 | 41.88 | -4.324 | 498 | .001 |
| 9. Noun overlap-adjacent | .311 | .231 | 3.459 | 498 | .001 |
| 10. Argument overlap-adjacent | .652 | .576 | 3.218 | 498 | .001 |
| 11. Stem overlap-adjacent | .371 | .271 | 4.163 | 498 | .001 |

| | | | | | |
|---|---|---|---|---|---|
| 12.Content word overlap-adjacent | .203 | .180 | 2.803 | 498 | .005 |
| 13. Noun overlap-all | .254 | .184 | 3.390 | 498 | .001 |
| 14. Argument overlap- all | .573 | .493 | 3.418 | 498 | .001 |
| 15. Stem overlap-all | .306 | .213 | 4.266 | 498 | .001 |
| 16. Content word overlap-all | .172 | .153 | 2.444 | 498 | .015 |
| 17. Logical connectives | 27.19 | 32.23 | -2.295 | 498 | .022 |
| 18. Temporal connectives | 5.31 | 15.52 | -7.094 | 498 | .001 |
| 19. Intentional verbs | 21.94 | 30.56 | -4.462 | 498 | .001 |
| 20. Modifiers per noun phrase | 0.69 | 0.77 | -3.386 | 498 | .001 |
| 21. Passive voice density | 0.40 | 2.60 | -4.408 | 498 | .001 |
| 22. Adverbial density | 16.12 | 28.26 | -6.537 | 498 | .001 |
| 23. Prepositional density | 60.05 | 92.34 | -9.790 | 498 | .001 |
| 24. Content word familiarity | 591.38 | 587.82 | 4.827 | 498 | .001 |
| 25. Content word concreteness | 413.10 | 395.88 | 4.837 | 498 | .001 |
| 26. Content word frequency | 2.514 | 2.453 | 2.888 | 498 | .004 |
| 27.Content word meaningfulness | 448 | 444 | 2.123 | 498 | .034 |
| 28. Coh-Metrix L2 readability | 30.70 | 28.72 | 2.435 | 498 | .015 |
| 29. Present tenses | 12.07 | 8.97 | 7.558 | 498 | .001 |
| 30. Past tenses | 0.73 | 2.46 | -6.894 | 498 | .001 |

**Table 1.** Results of independent samples T-tests between 4th and 5th grade essays

As can be seen in Table 1, 5th graders' essays were found to include a significantly higher number of words per sentence than their 4th grade counterparts. They were also characterised by significantly longer words in terms of average number of syllables per 100 words and letters per 100 words, while containing a higher percentage of unique words than their 4th grade counterparts, whose essays included a higher percentage of tokens per word family. Moreover, 5th grade written texts contained a slightly higher proportion of less frequent content words in the CELEX database along with a significantly lower number of concrete words than their 4th grade counterparts. They were also characterised by a significantly higher proportion of intentional verbs, all of which could have contributed to increased text complexity. In addition, data analysis showed that 5th graders' written essays texts contained a significantly higher proportion of all connectives, including logical and temporal ones. These connectives are related to a writer's deeper understanding of the relations in a text since connectives do not only clarify the relationships among ideas but also provide a clear structural pathway for the readers to follow. At the same time, 5th graders' written essays contained a significantly lower proportion of noun, argument, stem, and content word overlap across all sentences, as well as a higher proportion of passive sentences, adverbs, and modifiers per noun phrase.

On the other hand, 4th graders' essays were characterised by a significantly higher degree of text easability estimated through the Coh-Metrix Easability components. They were found to be syntactically simpler as they contained a significantly lower number of words per sentence and used simpler, more familiar syntactic structures that are less challenging to process. They also included a significantly higher percentage of concrete words along with increased referential cohesion. At the same time, the mean incidence of ideas contained in texts written by 4th grade students was much lower than that of ideas contained in 5th graders' texts, which might indicate that words pertaining to the same family were most often repeated by 4th grade students within each text. Moreover, lexical richness was lower for 4th grade essays, which might be taken to reflect the less diverse vocabulary present in the specific dataset. The adjusted mean frequency for content words was also significantly higher in 4th graders' written essays, another indication that a limited number of rare content words were present in such texts. Finally, the analysis revealed that essays written by 4th graders included a statistically significant higher percentage of noun overlap, argument overlap, stem overlap, and content word overlap between adjacent sentences along with a higher proportion of present tenses, whereas past tenses were more frequent in 5th graders' essays. The higher incidence of past tenses in more advanced texts could be interpreted as an indicator of text complexity.

## 4 Concluding remarks

The results of the study showed statistically significant differences between the linguistic features identified in the essays produced by EFL young learners at different levels of language competence. Overall the results of the present study are in agreement with the findings of the literature, and further support the notion that there seems to be a close relationship between EFL learners' writing development and their grade level. More specifically, linguistic development occurs in the later stages of writing development and is primarily related to producing more elaborate texts with more sophisticated words, more complex sentence structure, and fewer cohesive features as a function of grade level and L2 lexical growth. These results could provide valuable information to EFL teachers regarding the breadth of vocabulary knowledge their students need to have for enhancing their written essays. For instance, EFL teachers might wish to devote time to the revision and consolidation of words appearing in various frequency bands since such a vocabulary-based instruction could provide students with an increased vocabulary range to include in their pieces of writing. In addition, more in-depth linguistic features that could help EFL teachers take even more consistent and informed decisions when assessing texts written at different grade levels could include assessment of lexical density and lexical richness along with repetition of types and tokens.

Through the investigation of the significant relationships among a range of text variables, the present research attempted to provide evidence regarding the extent to which essays written at different grade levels can be distinguished from one another using a number of linguistic features related to propositional density, lexical sophistication, syntactic complexity, and cohesion. It also aspired to make a methodological contribution. Instead of examining a limited number of text variables independently, it made use of advanced text analysis software applications and investigated the impact of 135 text variables on text complexity. On the other hand, as with all studies, the implementation of this one presented a number of challenges and limitations that we hope we will overcome in future research. For instance, due to the fact that only a specific set of descriptive prompts included in their school coursebook was used, written texts inevitably belonged to a specific genre – expository. If the range of prompts increased, the generalisability of present results might be further strengthened. It would also be useful to extend the present analysis to texts at both lower and higher grade levels following a comparative corpus-based approach for evidence-based conclusions to be drawn from a much more extensive dataset.

## References

Berninger, V., Cartwright, A., Yates, C., Swanson, L., & Abbott, R. (1994). Developmental Skills Related to Writing and Reading Acquisition in the Intermediate Grades. *Reading and Writing, 6*(2),161–196.

Crossley, S. & McNamara, D. (2009). Computational Assessment of Lexical Differences in L1 and L2 Writing. *Journal of Second Language Writing, 18*(2),119–135.

Crossley, S., Salsbury, T., McNamara, D., & Jarvis, S. (2011). Predicting Lexical Proficiency in Language *Learner Texts using Computational Indices. Language Testing, 28*(4), 561–580.

Ferris, D. (2004). The “Grammar Correction” Debate In L2 Writing: Where Are We, and Where Do We Go From Here? (and What Do We Do in the Meantime …?). *Journal of Second Language Writing, 13*(1). 49–62.

Haswell, R. (2000). Documenting Improvement in College Writing: A Longitudinal Approach. *Written Communication, 17*(3), 307–352.

McNamara, D., Max, L., McCarthy, P., & Graesser. A. (2010). Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes, 47*(4), 292–330.

Shaaban, K. (2000). Assessment of Young Learners' Achievement in ESL Classes in the Lebanon. *Language, Culture and Curriculum, 13*(3), 306–317.

Silva, T. (1993). Toward an Understanding of the Distinct Nature of L2 Writing: The ESL Research and Its Implications. *TESOL Quarterly, 27*(4), 657–677.

Uhl-Chamot, A. & El-Dinary, P. (1999). Children's Learning Strategies in Language Immersion Classrooms. *The Modern Language Journal, 83*(3), 319–338.

Witte, S. & Faigley, L. (1981). Coherence, Cohesion, and Writing Quality. *College Composition and Communication 32*(2), 189–204.

# Cut Scores for Combined Constructs

**Beate Zeidler**, telc gGmbH, Germany

**Abstract:** This paper addresses the issue of setting cut scores for components of an examination where those components comprise several skills, while the performance level descriptors that are to be used are organised per individual skill. This means that the performance level descriptors do not actually describe the target performance. Yet the judges in the standard setting have to form an idea on which performance should pass, and which perfomance should fail. A method for helping judges to form such an idea was tried out in the context of an examination of German, the “fide Sprachnachweis”. Problems arising from such an undertaking are described, and the method that was used is presented and discussed.

## 1 Introduction

It is not uncommon for an examination to report compound results, eg. for “Speaking plus Listening”/”Reading plus Writing”. The German examination “fide Sprachnachweis” follows this pattern. The developers of the examination were thus faced with the task of defining cut scores for compound results, while at the same time adhering to the specification that these cut scores should relate to the Common European Framework of Reference for Languages (CEFR), which is largely organised by separate skills. The skills relating to interaction do describe reading and listening skills as interactive components but not, by definition, in their isolated form such as “listening as a member of an audience”. This task requires some thought relating to theoretical as well as practical matters.

## 2 Issues

The *Manual for relating language examinations to the Common European Framework* addresses the problem in Section 6.10.1 (“Standard setting across skills”), by discussing the compensatory and the conjunctive approach and recommending a compromise of both.

Following the compensatory approach, points are added across skills and a cut score is defined over the sum of points. The conjunctive approach requires a separate cut score for each skill, each of which have to be met in order to pass the examination. A compromise could mean that only some of the separate cut scores have to be met, or that on top of the overall cut score some “minimum performance” cut scores are defined for the separate skills.

While this covers the conceivable ways of dealing with the problem on a practical level, some questions remain open: does it make sense at all to have an overall cut score? And if so, how is it to be found?

### 2.1 Theoretical issues: admissibility of compound scores

From a construct point of view, some thought should be given to the question of what it is exactly that is to be measured, if the reference system does not have performance level descriptors for combined skills. For the purposes of this paper, this question of the qualitative meaning of the scores that are to be reported cannot be expanded on. It should however be kept in mind that the question exists.

From a measurement point of view, it should be checked whether the scale for which the cut score is to be set covers several dimensions, i.e. whether the skills are so far apart that a common cut score may not make sense. This can be done by factor analysis. Factor analysis was carried out on the basis of 220 candidate results from the pretesting, and, not surprisingly, did indeed indicate two dimensions in each of the components, which are however correlated. As this paper focusses on practical aspects, this issue cannot be fully elaborated here – suffice it to say that dimensionality was addressed by starting out with judging the skills separately, and then going on to define the cut score over the combination of skills.

### 2.2 Practical issues: conducting a standard setting

Standard setting was carried out with the help of a panel of 12 experts, who were experienced teachers of German and raters in other German examinations at the target levels, A1 to B1. Their task was to set cut scores for two examinations of the fide German suite, i.e. the A1/A2 exam and the A2/B1 exam, in each case for the for “Speaking plus Listening” and the ”Reading plus Writing” component, so that a total of eight cut scores had to be set.

Performance level descriptors were used, but mainly those relating to each individual skill. The gap between skills had to be bridged by tapping the experts’ minds for their ideas of “an A1 person”, “an A2 person” etc. The challenge was to help judges to externalise these ideas.

Before the standard setting a pretesting had been carried out with 220 candidates, so that empirical difficulty values could be used in the standard setting. The productive tasks had been rated by two trained raters.

The standard setting was carried out in two phases with several subprocesses, which are outlined below.

#### *2.2.1 Standard setting, phase 1: Looking at the receptive items (1 day)*

Separate cut scores for the receptive items were not required. However, a standard setting-type activity was conducted to establish a good idea of the characteristics and of the difficulty of the receptive items in each judge‘s mind. This served as input for the standard setting for the combined skills.

To help judges to focus on minimal compentence for each of the target levels, they were provided with the CEFR scales (“Hörverstehen allgemein”, “Ankündigungen, Durchsagen und Anweisungen verstehen”, “Muttersprachliche Gesprächspartner verstehen”, “Leseverstehen allgemein”, and “Zur Orientierung lesen”, taken from Europarat 2001) for adjacent levels and asked to pinpoint the sort of performance that would distinguish a candidate who had just made the transition to the next (target) level. Then four items were discussed as to whether they could be expected to be solved by a minimally competent person (MCP), to see whether the panel had formed a uniform idea of minimal competence for each level.

The main activity was inspired by Wright & Grosse (1993) and consisted of judging for each item whether it was an item that could be answered correctly by an MCP at Levels A1, A2 and B1 respectively, and whether it was an item at the borderline between levels. It was thus required to go through the items three times (once for each level). Mean difficulty of these borderline items was calculated to define a provisional cut score for the receptive skills.

Judges‘ perception of the items could be shown to be consistent with the empirical item difficulties, so that the activity could probably regarded as successful in that judges formed a realistic idea of the items‘ difficulty (Kaftandjieva’s (2010, p. 57) Misplacement Index in all cases above 0.64=at least acceptable, in most cases good).

### *2.2.2 Standard setting, phase 2: Looking at receptive plus productive skills (2 days)*

For this phase candidate dossiers had been prepared along the lines suggested by Sweeney & Ferdous (2007). This method was deemed to be promising, because it allows the integration of answers to multiple-choice items and productive performance, and because it requires only one pass through candidate performances.

20 candidates for Reading+Writing A1-A2 and Reading+Writing A2-B1 respectively, 4 candidates for Listening+Speaking A1-A2 and 5 candidates for Listening+Speaking A2-B1 had been chosen to represent a good spread over abilities, and to also include some predictably problematic cases, such as candidates writing not to the topic.

Each dossier showed a list of receptive items, arranged in ascending order of difficulty and including a short description of the item, its content area and information on whether the specific candidate whose dossier it was had given a correct answer or not. The number of correct answers was also given. Dossiers for Reading+Writing contained information on the reading items as well as a reproduction of the text written by the candidate, dossiers for Speaking+Listening contained information on the listening items. Speaking performances were shown on video.

In addition, each judge had a list of all candidates for the subtest under scrutiny, arranged in ascending order of their raw scores. These were referenced to the dossiers by way of the candidate numbers, and contained a summary on the percentage score reached for each skill, i.e. for Reading+Writing: reading items, writing-related tasks and productive writing, and for Listening+Speaking: listening items and speaking. Their first task was to decide, based on this information and on the dossiers, which CEFR level each candidate had reached. After this had been done for all candidates, the second task was to decide which was the weakest candidate who would attain the target level, and which was the top failing candidate. Finally, they had to decide on a cut score which must be between these two candidates’ scores. Judges were free to choose the mean, or any other score in the range between the two scores. These tasks were carried out for two cut scores at a time, i.e. four times in all, by each judge independently.

All proposed cut scores were presented to the group when one part of the examination had been dealt with. For Speaking+Listening A1/A2, A1 cut scores ranged from 38 to 45 with a mean of 40.22 and A2 cut scores ranged from 55 to 65 with a mean of 58.44, for the A2/B1 exam the A2 range was 58 to 60 with a mean of 58.5, and the B1 range 78 to 80 with a mean of 79.75.

For Reading+Writing A1/A2, A1 cut scores ranged from 40 to 42 with a mean of 40.36 and A2 cut scores ranged from 55 to 60 with a mean of 57.55, for the A2/B1 exam the A2 range was 42 to 57 with a mean of 49.6, and the B1 range 65 to 76 with a mean of 70.2. It was apparently easiest to define cut scores for Reading+Writing A1, Listening+Speaking A2 and B1 while there was the highest divergence for Reading+Writing in the A2/B1 examination.

These results, as well as the positioning of each judge’s cut score, were presented graphically to the judges, and formed the basis for a on whether candidates near the cut scores were assigned a plausible level, and what was a plausible level assignment in the case of diverging abilites. Judges then had the opportunity to modify their judgements.

### *2.2.3 Standard setting, results*

For Speaking+Listening, the process described above led to a consensus on the following cut scores: A1/A2 exam, A1: 40, A2: 58; A2/B1 exam: A2: 59, B1: 80. For Reading+Writing, in the A1/A2 exam, A1: 40, A2: 58; in the A2/B1 exam, A2: 45, B1: 68. While the modifications in the second round were insignificant for Listening+Speaking and for Reading+Writing A1/A2, cut scores for Reading+Writing A2/B1 were perceptibly lowered.

Discussion also showed the need for the introduction of an additional rule, namely, a minimum score for each skill that has to be passed in order to consider the candidate’s performance as a possible overall pass. This was fixed at 6 percent, which means that for speaking at least A1 has to be reached in all content criteria, or a minimum competence in the language criteria. For writing, it is possible to pass with 0 points for productive writing, but in this case at least 24 percent of the possible points for writing-related tasks have to be reached. In the receptive skills, at least one item has to be answered correctly. The approach taken is thus not wholly compensatory, in that it is not possible to attain a level for one of the exams without any performance in either one of the skills.

## 3 Limitations

There are of course some limitations to this standard setting, due to practical issues. The number of candidate samples was limited, especially in the case of the speaking performances, mainly because time did not permit to watch and discuss more videos. It would also have been desirable to invite more judges. However, a consensus could be reached which all participants described as a good basis for work with the examination. Judges were asked whether they felt confidence in their common result (“Wie stark vertrauen Sie den von der Gruppe ermittelten Grenzwerten?”), with four options: strongly confident, somewhat confident, less confident, not confident. Most answers were “strongly confident”, with the exception of three “somewhat confident” for Speaking+Listening, and four “somewhat confident” for Reading+Writing.

## 4 Conclusion

It was attempted to find a plausible way of fixing a cut score for a combination of skills, without being able to recur to performance level descriptors that would describe the exact target performance. In order to do so, judges had to be given a setting that would allow them to exercise their judgement as to “A1”, “A2”, “B1” competence, which many language teachers feel able to do, while perhaps not being able to describe target competence theoretically or abstractly. The judges who were all language teachers were presented with evidence of competence in a way that was as transparent as possible. A plausible result was reached in which all participants in the standard setting expressed confidence.

## References

Europarat. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheidt.

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests. A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem: EALTA.

Sweeney, K. P. & Ferdous, A. (2007). *Variations of the ‘body of work’ standard setting method*. Paper presented at the Annual meeting of the National Council on Measurement in Education, Chicago.

Wright B. D. & Grosse M. (1993). How to set standards, *Rasch Measurement Transactions, 7*(3), 315.

# Towards a Scale of Academic Language Proficiency

**Stuart Shaw**, Cambridge International Examinations, UK
**Helen Imam**, Cambridge International Examinations, UK

**Abstract:** With increasing use of the Common European Framework of Reference for Languages (CEFR) in international school contexts and with the blurring of first, second or foreign language distinctions, we believe that there is a potential need for a supplementary scale of academic language proficiency. Many educators often tend to focus on the listening, speaking, reading and writing scales of the CEFR or on the global scale, which only touches on academic contexts as it has to encapsulate other contexts such as a social or foreign language. The CEFR provides a wealth of specialised scales - for example, the text processing scale, the pragmatic scales (thematic development, propositional precision, coherence and cohesion), the strategic scale of compensating, as well as tables that combine scales which draw together "relevant qualitative factors for production" (Council of Europe, 2009, p. 149). However, aspects of academic language are found across various scales, which can make it hard to locate and apply them to school contexts. Furthermore, it is often assumed that academic language proficiency features in the upper parts (higher proficiency) of the CEFR scale, where there are high expectations for foreign language learners that would not always be met even by learners for whom English is a first language. Conversely, the lower end of the CEFR scales might not always capture the academic language that learners may be developing in their early stages. In this presentation we propose an academic language proficiency scale that would draw together aspects of academic language ability found in other scales and, if needed, could add new skills not currently covered by the CEFR. It might even be possible to provide a memorable quality name for each level, with descriptors of academic language proficiency either for individual subjects or in general.

*Council of Europe. (2009). Relating language examinations to the Common European framework of reference for languages: learning, teaching, assessment (CEFR): a manual. Strasbourg: Language Policy Division, Council of Europe.*

## 1 Introduction

English is a global language – a lingua franca or lingua mondo. It is also the medium of instruction and assessment for Cambridge International Examinations (hereafter Cambridge) programmes of learning and assessment. Cambridge develops and provides programmes of learning and assessments worldwide in a wide range of subjects. These programmes of learning are delivered by schools all over the world in a variety of multilingual and educational contexts, and increasingly in bilingual education contexts. One key function of these programmes is to prepare students whose first language (L1) is not necessarily English as candidates for international high-stakes assessments.

This international context poses both a potential threat to, and an opportunity for, language development. The international quest for English and for an English-medium education can cause anxieties about achievement through the L2, as well as about the maintenance of L1s. An alternative to wholly English-medium education is bilingual education, in which two languages are used within the curriculum as mediums of instruction for non-language content subjects. Learning some content subjects (such as science and history) through an L2, and other content subjects through L1s, can create authentic language environments as students are immersed in and have to use both languages for communication about meaningful content. Bilingual education is a fast-developing practice that is becoming an increasingly widespread direction of language learning in schools (Mehisto & Genesee, 2015).

In the research described here we attempt to link Cambridge international assessments to the Common European Framework of Reference for Languages (CEFR) and provide estimates of minimum CEFR language levels in order to aid teachers in the preparation of their students for Cambridge international programmes of learning and summative assessment delivered through the medium of English. We also reveal how the research outcomes could inform the construction of an academic language scale.

## 2 Context: the Cambridge international curriculum

Students preparing for Cambridge international school qualifications do so in very diverse linguistic and educational contexts. Some schools follow an entire curriculum in English, while others teach only a few subjects in English. Cambridge international programmes and qualifications are often used for the English-medium strand of a bilingual education programme and are taken alongside qualifications from students' own (non-English) national curriculum.

The Cambridge international curriculum is a continuum starting at the age of 5 and running through to the age of 18 or 19. The programmes are progressive, embodying the same commitment to the acquisition and exercise of higher order skills, deep understanding and confidence in applying learning at every stage. There are four stages in Cambridge international programmes and qualifications. Cambridge Primary and Secondary 1 include diagnostic feedback, while Cambridge Secondary 2 and Cambridge Advanced include the International General Certificate of Secondary Education (IGCSE) and the International Advanced Subsidiary (AS)/Advanced (A) level, respectively. Schools following the programme can offer all four stages, or one or two stages only. Each stage is designed to build on learners' development in the previous stage. A key juncture for all these students takes place around the age of 16, when students typically enter for the Cambridge IGCSE qualification in a range of subjects.

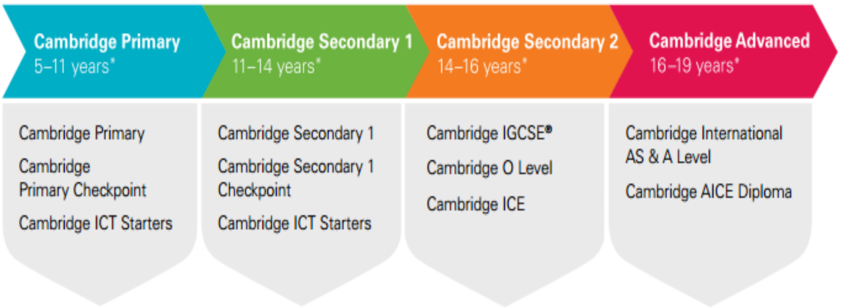


**Figure 1.** The Cambridge curriculum

## 3 Researching the linguistic demands of content assessments

Shaw & Imam (2013, referring to Shaw 2011, 2012) describe an analysis of the linguistic demands in Cambridge International General Certificate of Secondary Education (IGCSE) history, biology and geography assessment instruments. These subjects were chosen because their assessments use a range of question types, because they are popular subjects, and because the nature of the subjects provides a reasonable basis from which to consider the generalisability of the findings to a range of other subjects; for example, the humanities, sciences and social sciences. The study sought to address the following question: What English language skills are needed to understand typical Cambridge IGCSE assessments and succeed in them? Data for analysis included syllabuses, question papers, mark schemes and candidates' performances.

In addition to identifying the types of academic language skills required by three Cambridge IGCSE content-based assessments, the Shaw and Imam study (2013) also evaluated the linguistic inputs and outputs of these IGCSE subjects against a common European scale of language proficiency – the CEFR.

Estimates of the CEFR language levels needed by students enable teachers to better prepare their students for programmes and assessments such as the Cambridge IGCSE. The reason for this is that international teachers of content often want to know the level of English that their students need in order to have a chance of success in an IGCSE content subject. Also, the successful attainment of IGCSE non-language qualifications gives added value to a bilingual education programme and indicates that a student has sufficient English language proficiency to be able to cope with academic content being taught through English.

By analysing test instruments as well as candidate responses, Shaw and Imam (2013) attempted to identify general minimum English language levels needed to access IGCSE content.[8] Their research suggests that students for whom English is not an L1 do not appear to be disadvantaged at IGCSE level in terms of their English language skills if they are at least B2 independent users of English, according to the CEFR levels.

The researchers concluded that an average language proficiency level of B2[9] on the CEFR (Council of Europe, 2001a) is useful to access typical IGCSE exams, and CEFR level of C1[10] could provide an added advantage of linguistic resources to be able to develop arguments needed for high grades for humanities subjects such as history and geography. Other research corroborates the B2/C1 finding (Imam, 2010). There is evidence to suggest that CEFR Level B2 could represent a critical CALP level for the 15-16 age group.

[8] The purpose of the Shaw and Imam (2013) study was not to benchmark the English language demands of geography, history and biology IGCSEs to the CEFR. The primary focus was on the linguistic demands of the assessment inputs (syllabus, question papers and mark schemes) and outputs (candidate performances). Suggested links to the CEFR at a global level are therefore tentative.

[9] An independent user B2 can "understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation.… interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party.… produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and Independent disadvantages of various options" (Council of Europe, p. 24).

[10] A proficient user C1 can "understand a wide range of demanding, longer texts, and recognise implicit meaning … express themselves fluently and spontaneously without much obvious searching for expressions.… use language flexibly and effectively for social, academic and professional purposes.… produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices" (Council of Europe, p. 24).

This finding is consistent with on-the-ground practice in some countries. For example, in Dutch bilingual schools CEFR B1/B2 (depending on the type of secondary education) is the required English standard at the end of the third grade of secondary education (which would be around the age of 15). In Colombia, based on anecdotal feedback from some schools, CEFR B1 is thought necessary for starting an IGCSE programme and CEFR B2 is thought necessary for the end of an IGCSE programme and for accessing IGCSE exams. There is therefore evidence that CEFR Level B2 could represent a critical level for this age group (around the age of 15–16) in educational contexts taking English-medium L2 assessments.

Interestingly, the Shaw and Imam study (2013) found that there were few problems with candidates misunderstanding and using subject-specific vocabulary. Such vocabulary, due to its low frequency, might be identified by certain vocabulary databases to be as high as CEFR C2 level. This suggests that even if a candidate is generally at a lower CEFR level (for example, B2), the teaching and learning of such low-frequency specialist vocabulary is part and parcel of the teaching and learning of the concepts of a content subject. This also negates any potential argument for the strict observance of CEFR levels for using vocabulary when writing content curricula and assessments.

It should be noted, however, that a minimal level of language proficiency may be necessary but not sufficient for success in a content-based exam as other factors are also relevant, such as knowledge of the syllabus content or, as Srole (1997) indicates, cultural exposure: “This fusion of language and content requires students to understand non-history content and cultural references that linguistically and culturally diverse students do not yet possess” (p. 105).

## 4 Developing a supplementary CEFR scale of academic language proficiency

In the Shaw and Imam study (2013) the CEFR served as a widely recognised tool to indicate useful proficiency levels for certain content assessments. It was also a useful tool for capturing relevant academic language skills, although not all academic language skills found in Cambridge international content assessments were found in the CEFR. The CEFR was designed with European adult foreign language learners in mind but was intended to be adaptable to individual contexts. The Shaw and Imam (2013) study involved drawing together aspects of the CEFR relevant to academic language proficiency in different subjects. We would like to take this work further and help to develop a supplementary scale focusing on academic language proficiency, with descriptors for each CEFR level. This would have the ultimate goal of helping school educators plot the progress of their students in the key academic language needed to achieve in content subjects. Such a scale, with detailed descriptors of academic language, might even be found to be applicable to academic school contexts and to students in general, whether English is a first or additional language. We believe targeting this wider student group would not be inconsistent with the philosophy of the CEFR, which acknowledges that a learner’s cognitive processes and skills develop through engagement with the communicative tasks that arise in social interaction.

With increasing use of the CEFR in international school contexts and with the blurring of first, second or foreign language distinctions, we believe that there is a potential need for a supplementary scale of academic language proficiency. Many educators often tend to focus on the listening, speaking, reading and writing scales of the CEFR or on the global scale, which only touches on academic contexts as it has to encapsulate other contexts such as a social or foreign language. The CEFR provides a wealth of specialised scales (Council of Europe, 2001b)[11] – for example, the text processing scale, the pragmatic scales (thematic development, propositional precision, coherence and cohesion), the strategic scale of compensating, as well as tables that combine scales – for example, CEFR's table A5, which draws together "relevant qualitative factors for production" (Council of Europe, 2009, p. 149). However, aspects of academic language are found across various scales, which can make it hard to locate and apply them to school contexts. Furthermore, it is often assumed that academic language proficiency features in the upper parts (higher proficiency) of the CEFR scale, where there are high expectations for foreign language learners that would not always be met even by learners for whom English is an L1. Conversely, the lower end of the CEFR scales might not always capture the academic language that learners may be developing in their early stages. The proposed new scale would draw together aspects of academic language ability found in other scales and, if needed, could add new skills not currently covered by the CEFR. It might even be possible to provide a memorable quality name for each level, with descriptors of academic language proficiency either for individual subjects or in general. For example, our research into Cambridge IGCSE history, involving reading (sources) and writing, led to the following beginnings of a scale (See Table 1)

| **CEFR history level** | **Quality** | **Descriptor** | **CEFR scales** |
|---|---|---|---|
| CEFR: C2<br>history: bonus marks | 'Evaluate & create' | *CEFR*<br>Coherent and cohesive<br>Reconstructs arguments from different sources<br>Clear, complex, logical<br>Smooth substitution for specialist words<br>IGCSE history mark scheme<br>Bonus marks: evaluation of sources | *Pragmatic*<br>Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices<br>*Text processing*<br>Can summarise information from different sources, reconstructing arguments and accounts in a coherent presentation of the overall result<br>*Overall written production*<br>Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points<br>*Reading for information & argument*<br>No descriptor available<br>*Strategic*<br>Can substitute an equivalent term for a word he/she can't recall so smoothly that it is scarcely noticeable<br>*Socio-linguistic*<br>Appreciates fully the socio-linguistic and sociocultural implications of language used by native speakers and can react accordingly |

[11] For a compilation of all the scales from chapters 3, 4 and 5 of the CEFR, see the structured overview of all CEFR scales (Council of Europe, 2001b).

| | | | |
|---|---|---|---|
| CEFR: C1<br>history: levels 4 & 5 | ‘Structure & justify’ | CEFR<br>Clear and well-structured<br>Effective<br>Summarises long texts<br>Expanding and supporting details<br>Conclusion<br>Reformulates without interrupting flow<br>IGCSE history mark scheme:<br>Level 4: implicit match or mismatch:<br>implicit understanding of one side of an argument<br>↓<br>Level 5: implicit match and mismatch:<br>implicit understanding of both sides of an argument | *Pragmatic*<br>Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices<br>Can give elaborate descriptions and narratives, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion<br>*Text processing*<br>Can summarise long, demanding texts<br>*Reading for information & argument*<br>Can understand in detail a wide range of lengthy, complex texts … identifying finer points of detail including attitudes and implied as well as stated opinions<br>*Overall written production*<br>Can write clear, well-structured texts on complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion<br>*Strategic*<br>Can backtrack when he/she encounters a difficulty and reformulate what he/she wants to say without fully interrupting the flow of speech<br>*Socio-linguistic*<br>Can use language flexibly and effectively for social purposes, including emotional, allusive and joking usage<br>*Reports and essays*<br>Clear, well-structured expositions |
| CEFR: B2<br>history: levels 3 & 4 | ‘Detail & analyse’ | CEFR:<br>Clear, limited cohesion<br>Relevant expanding and supporting detail<br>Comments on contrasting points of view and main themes<br>Synthesise and evaluates number of sources<br>Paraphrases<br>Appropriate<br>IGCSE history mark scheme:<br>Level 3: surface match or mismatch:<br>Explicit understanding of one side of an argument<br>↓<br>Level 4: surface match and mismatch:<br>Explicit understanding of both sides of an argument | *Pragmatic*<br>Can develop a clear description or narrative, expanding and supporting his/her main points with relevant supporting detail and examples<br>Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some “jumpiness” in a long contribution.<br>*Text processing*<br>Can summarise a wide range of factual and imaginative texts, commenting on and discussing contrasting points of view and the main themes<br>Can summarise extracts from news items, interviews or documentaries containing opinions, argument and discussion<br>Can summarise the plot and sequence of events in a film or play<br>*Reading for information & argument*<br>Can obtain information, ideas and opinions from highly specialised sources within his/her field...<br>*Overall written production*<br>Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources<br>*Strategic*<br>Can use circumlocution and paraphrase to cover gaps in vocabulary and structure<br>Can make a note of favourite mistakes and consciously monitor speech for them<br>*Socio-linguistic* |

| | | | |
|---|---|---|---|
| | | | Can express him or herself appropriately in situations and avoid crass errors of formulation<br>*General linguistic range*<br>Upper B2: broad range of language to express clearly<br>Lower B2: express viewpoints and develop arguments |
| CEFR: B1<br>history: levels 1 & 2 | 'Describe' | CEFR<br>Linear description<br>Collate short pieces<br>Simple substitution<br>IGCSE history mark scheme:<br>Level 1: writes about the sources but no comparison/no valid use of sources<br>Level 2: compares details but no attitudes compared | *Pragmatic*<br>Can link a series of shorter, discrete simple elements in order to reasonably fluently relate a straightforward narrative or description as a linear sequence of points<br>*Text processing*<br>Can collate short pieces of information from several sources and summarise them for somebody else.<br>Can paraphrase short written passages in a simple fashion, using the original text wording and ordering<br>*Reading for information & argument*<br>Can identify the main conclusions in clearly signalled argumentative texts. Can recognise the line of argument … though not necessarily in detail<br>*Overall written production*<br>Can write straightforward connected texts on a range of familiar subjects within his/her field of interest, by linking a series of shorter discrete elements into a linear sequence<br>*Strategic*<br>Can use a simple word meaning something similar to the concept he/she wants to convey and invites "correction"..<br>Can start again using a different tactic when communication breaks down<br>*Socio-linguistic*<br>No descriptor available |
| A2 | 'Link' | CEFR<br>Links groups of words with simple connectors<br>Picks out key phrases<br>Series of simple phrases and sentences | *Pragmatic*<br>Can link groups of words with simple connectors like 'and', 'but' and 'because'<br>*Text processing*<br>Can pick out and reproduce key words and phrases or short sentences from a short text within the learner's limited competence and experience<br>*Reading for information & argument*<br>Can identify specific information in simpler written material … such as letters, brochures and short newspaper articles describing events<br>*Overall written production*<br>Can copy out short texts in printed or clearly handwritten format<br>Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'<br>*Strategic*<br>No descriptor available<br>*Socio-linguistic*<br>No descriptor available |
| A1 | 'Identify' | CEFR<br>Copies single words<br>Links words with linear | *Pragmatic*<br>Can link words or groups of words with very basic linear connectors like 'and' or 'then' |

| | | | |
|---|---|---|---|
| | | connectors<br>Simple isolated phrases and sentences | *Text processing*<br>Can copy out single words and short texts presented in standard printed format<br>*Reading for information & argument*<br>Can get an idea of the content of simple information and short descriptions, especially if there is visual support<br>*Overall written production*<br>Can write simple isolated phrases and sentences<br>*Strategic*<br>No descriptor available<br>*Socio-linguistic*<br>No descriptor available |

**Table 1.** Tentative academic language proficiency scale: for example, IGCSE history

Our work on this scale is in its initial stages. It is not a simple task to compare descriptors between scales. For example, the word evaluate in the History mark scheme attracts the highest number of marks, whereas the word evaluate appears only in the upper B2 level descriptors of the CEFR. This means one cannot rely simply on a surface comparison of words in the two scales as they might mean different things when unpacked.

However, Table 1 illustrates that a key CEFR level for IGCSE history could be B2, which moves language beyond the descriptive realm (B1) into the analytic realm. A crucial jump in the history mark scheme is from explicit understanding to implicit understanding of texts. Understanding implied opinions appear in the CEFR from Level C1. However, there may not be a consistent correspondence between language and history at the individual level. Clearly, a student who is at CEFR B1 level or lower, using simple, descriptive language, would not have the language to be able to access, analyse and evaluate source material. Another student may have the sophistication of language at CEFR C2 level but may not have sufficient cognitive ability or historical knowledge or exam technique to evaluate history source material and gain marks at the highest level of the mark scheme. Conversely, a student with less sophisticated language at CEFR C1 or B2 level still may be able to grasp the content and effectively communicate their evaluation to examiners. This can also be seen in the Table 2. Imam (2010) suggests that the IGCSE history grade descriptor references to language could correspond to more than one CEFR language level. For brevity, only IGCSE history grades A and C are shown here to illustrate the point.

| **History syllabus CEFR:**<br>**Grade** | **Grade descriptor** | **CEFR scale and level** |
|---|---|---|
| Grade A | Recall, select and deploy relevant historical knowledge accurately to support a **coherent and logical argument** | *General linguistic range*<br>Upper B2: can express him/herself clearly and without much sign of having to restrict what he/she wants to say. Has sufficient range of language to be able to...**develop** |

| | | |
|---|---|---|
| | | **arguments**...using some complex sentence forms.<br>C1: ...broad range of language to express him/herself clearly, without having to restrict what he/she wants to say.<br>C2: can exploit a comprehensive and reliable mastery of a very wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity. No signs of having to restrict what he/she wants to say.<br>*Writing report and essays*<br>Lower B2: develops **an argument**, giving reasons in support of or against a particular point of view … can synthesise information and arguments from a number of sources.<br>Upper B2: develops an argument systematically with appropriate highlighting of significant point and relevant supporting detail. Can evaluate different ideas or solutions to a problem.<br>C1: clear, well-structured expositions of complex subjects, underlining the relevant salient issues. Can expand and support points of view at some length with subsidiary points, reasons and relevant examples.<br>C2: smoothly flowing, complex … essays which present a case, or give critical appreciation of proposal or literary works … appropriate and effective logical structure which helps the reader to find significant points. |
| Grade C | Recall, select and deploy relevant historical knowledge **in support of a logical argument** | *General linguistic range*<br>Mid B2: expresses viewpoints and **develops arguments** … using some complex sentence forms to do so.<br>*Writing report and essays*<br>Lower B2: **develops an argument** … giving reasons … can synthesise information and arguments from a number of sources. |
| Grade A | Communicate in a **clear and coherent** manner using **appropriate historical terminology** | *General linguistic range*<br>Upper B2: can express him/herself **clearly** … without having to restrict what he/she wants to say<br>*Vocabulary range*<br>B2: **good range of vocabulary for matters connected to his/her field** … lexical gaps cause circumlocution.<br>C1: broad lexical repertoire allowing gaps to be readily overcome with circumlocutions<br>C2: very broad lexical repertoire … shows awareness of connotative levels of meaning.<br>*Vocabulary control*<br>B2: **lexical accuracy generally high** … some incorrect word choice without hindering communication.<br>C1: occasional minor slips.<br>C2: consistently correct and appropriate use of vocabulary.<br>*Overall written production*<br>C1: **clear, well-structured** … expanding and supporting points of view … appropriate conclusion<br>C2: clear, smoothly flowing, complex texts … appropriate and effective style … logical structure that helps the reader to find significant points.<br>*Coherence*<br>Lower B2: limited number of cohesive devices to link utterances into **clear, coherent discourse**, though there |

| | | |
|---|---|---|
| | | may be some jumpiness in a long contribution<br>Upper B2: variety of linking words efficiently to mark clearly the relationship between ideas<br>C1: clear, smoothly flowing, well-structured … showing controlled use of organisational patterns, connectors and cohesive devices.<br>C2: coherent and cohesive text making full and appropriate use of organisational patterns and a wide range of cohesive devices. |
| Grade C | Communicate in a **clear and coherent** form using **appropriate historical terminology** | *General linguistic range*<br>Lower B2: sufficient range of language to give **clear** descriptions, express viewpoints and develop arguments … some complex sentence forms<br>*Vocabulary range*<br>B2: **good range of vocabulary for matters connected to his/her field** … lexical gaps cause circumlocution<br>*Vocabulary control*<br>B2: **lexical accuracy high** … some incorrect word choice without hindering communication<br>*Overall written production*<br>B2: **clear, detailed**<br>*Coherence*<br>As Grade A |
| Grade A | Interpret and evaluate a **wide range** of historical sources and their use as evidence; **identify precisely the limitations** of particular sources; **compare and contrast** a range of sources and draw **clear, logical conclusions** | *Processing text*<br>B2: can summarise **wide range of** … **texts** … commenting on and discussing **contrasting** points of view and main themes<br>C1: can summarise long, demanding texts<br>C2: can summarise information from different sources, reconstructing arguments and accounts in a coherent presentation<br>*Overall written production*<br>Upper B2: **clear, detailed** … **synthesising and evaluating** information and arguments from a **number of sources**<br>C1: clear, well-structured expositions of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion<br>C2: clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points. |
| Grade C | Interpret and evaluate historical sources and their use as evidence; **indicate the limitations** of particular sources; compare and contrast a range of sources and draw **coherent conclusions** | *Processing text*<br>Same as Grade A<br>*Overall written production*<br>Same as Grade A |

**Table 2.** Comparison of IGCSE history grade descriptors with CEFR level descriptors

The variation shown here was borne out in the exam performance data of individual candidates who had achieved higher or lower grades in IGCSE English as a second language compared to their grades in IGCSE history, although overall there was a general language dependency found in IGCSE history.

## 5 Conclusion

Our work on this academic language proficiency scale is in its initial stages and needs further development. We are also gathering information about other proficiency scales besides the CEFR, such as WIDA's English Language Development Standards (USA), the cognitive academic language learning approach (USA), the National Association for Language Development in the Curriculum, English as a second/additional language formative assessment descriptors (UK), and FörMig key-stage descriptors for German as a second language (Germany). We are also aware that the Council of Europe and the European Centre for Modern Languages (ECML) are engaged in a related development that is much broader in scope than the focus of our research, possibly involving plurilingual and intercultural competences.

The development of an academic language proficiency scale is complex and multidimensional, as it inevitably introduces a range of factors: the student's cognitive stage, general language proficiency, and the processes and skills involved in mastering the specific curricular objectives of each subject area, as well as the processes and skills involved in learning in general. It cannot be assumed that these processes and skills are the same across countries or cultures, given possibly different educational traditions and modes of discourse.

Finally, the exact purpose of such an academic language proficiency scale needs to be determined – whether it would be a series of descriptors of language use in content classrooms and assessments, whether it would target specific programmes, subjects and students' ages (such as IGCSE history typically at the age of 16) or whether it would be more generic. It also needs to be determined whether it would be a tool supporting intervention in language use in school classrooms. The latter implies that a solid understanding of how learning happens, and what role language plays in the process must be at its centre.

## References

Council of Europe. (2001a). *Common European framework of reference for languages: learning, teaching, assessmen*t. Cambridge: Cambridge University Press.

Council of Europe. (2001b). *Common European framework of reference for languages: learning, teaching, assessment. Structured overview of all CEFR scales*. Retrieved from http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/Overview_CEFRscales_EN.pdf

Council of Europe. (2009). *Relating language examinations to the Common European framework of reference for languages: learning, teaching, assessment (CEFR): a manual*. Strasbourg: Language Policy Division, Council of Europe.

Imam, H. (2010). *Calpability: achieving in content through language: teacher perceptions, examiner expectations and student performance in IGCSE history*. Unpublished MA dissertation. Anglia Ruskin University.

Mehisto, P. & Genesee, F. (Eds.) (2015). *Building bilingual education systems: forces, mechanisms and counterweights*. The Cambridge Education Research series. Cambridge: Cambridge University Press.

Shaw, S. D. (2011). *Investigating the relationship between performance in language assessment and other, non-language IGCSE subjects. Phase 1: analysis of question papers and mark schemes. Phase 2: analysis of candidate output*. Cambridge: Cambridge International Examinations.

Shaw, S. D. (2012). International assessment through the medium of English: analysing the language skills required. *Research Matters, 13*, 2–10.

Shaw, S. D. & Imam, H. C. (2013). Assessment of international students through the medium of English: ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly, 10*, 452–475.

Srole, C. (1997). Pedagogical responses from content faculty: teaching content and language in history. In M. A. Snow & D. M. Brinton (Eds.), *The content-based classroom: Perspective on integrating language and content* (pp. 104–116). Harlow: Longman.

# Developing a Japanese Language Test for a Multilingual Online Assessment System: Towards an Action-oriented Approach to Japanese Instruction in Europe

**Tomoko Higashi**, Grenoble Alpes University, Lidilem, France
**Chieko Shirota**, Grenoble Alpes University, France
**Michiko Nagata**, Grenoble Alpes University, France

**Abstract:** In order to respond to the need for transparent and common assessment criteria for European students learning the Japanese language, we have undertaken to develop a CEFR-based Japanese test in the multilingual assessment system "SELF" as a part of the Innvalangues project (Université Grenoble Alpes). In this paper, after an overview of SELF, we will present our approach and point out particular difficulties in the development of a test in a non-European language as a part of a multilingual common framework. We will highlight the risk of bias relative to sociocultural knowledge in a language test when the target language is socio-culturally distant from the learner's language and present our reflection with examples of our tasks and items.

## 1 Introduction

The number of learners of Japanese as a foreign language has been steadily increasing for the last decade in France (The Japan Foundation 2003; 2017). In 2015, France is the first country in Europe regarding the number of learners of Japanese, with more than 20,000 learners (The Japan Foundation, 2003; 2017). Taking into account the increase in the number of learners and their diversity, as well as the development of international mobility, there is a clear need for transparent and common assessment criteria for European students learning Japanese. In these situations, undertaking to develop a CEFR-based Japanese test in the multilingual assessment system "SELF" would contribute to fill this gap. In this paper, after an overview of SELF, we will point out particular difficulties the development of a test in a non-European language as a part of a multilingual common framework. In this study, we will highlight the risk of bias relative to sociocultural knowledge in a language test when the target language is socio-culturally distant from the learner's language and present our reflection with examples of our tasks and items.

## 2 Overview of multilingual online test SELF

SELF means "Système d'Évaluation en Langues à visée Formative" (Assessment system of foreign languages with formative aim), which is a part of the "Innovalangues" project, winner of a National Research Grant, IDEFI, ("Initiative of Excellence for innovative formation") supported by Grenoble University. SELF can be used as placement test, but also as a proficiency test with a formative and diagnostic aim. It is an online-based and adaptive test, assessing three abilities: listening, reading and short writing. The general result and three separated results are shown, allowing each learner to become aware of their strong and weak points. It takes into account partial competence, suggested by the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001) (CEFR 6.1.3.4), and the general result advises on the optimal level course to attend.

SELF is a multilingual assessment system, available in Italian and English (for A1 to C1 level), and Chinese (A1 to B1). Furthermore, Japanese, Spanish and French as foreign language tests are in the process of being developed. The Japanese test is scheduled to come into service in September 2017 (A1 to B1).

## 3 Developing the Japanese test in the multilingual framework

### 3.1 Process in the test development cycle

SELF is based on a common methodological approach for these languages. The test development is based on a model of qualitative and quantitative validation, represented by an iterative process of successive steps (see Figure 1).

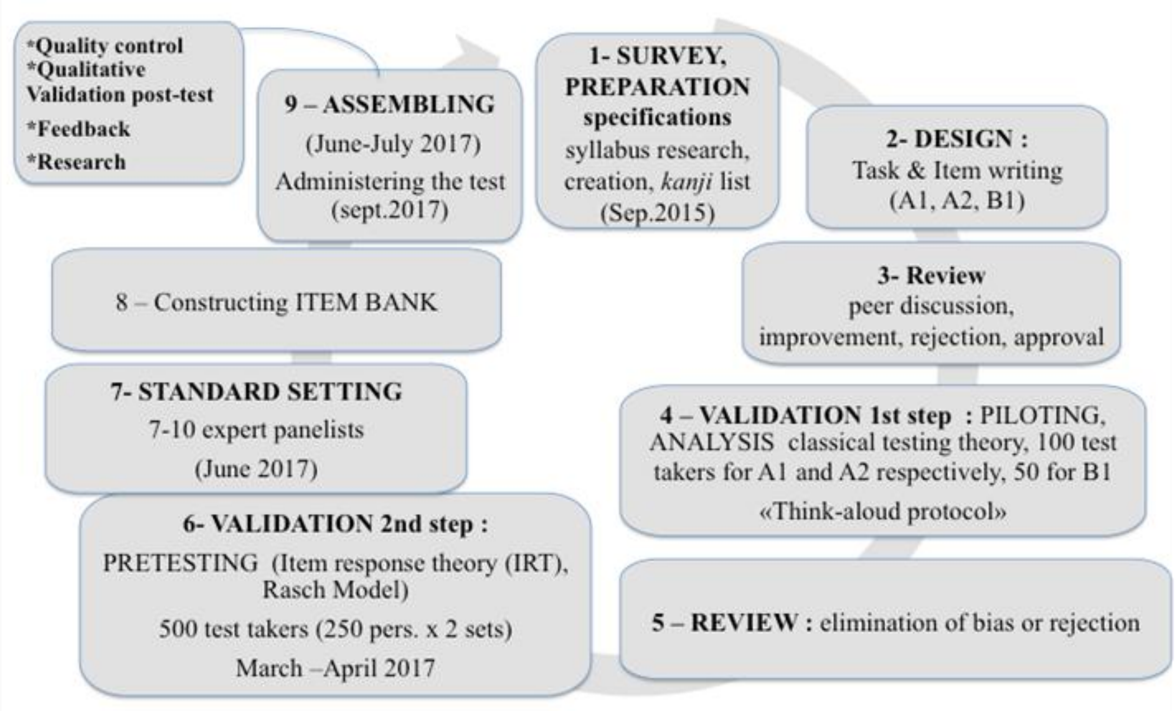

*Quality control
*Qualitative Validation post-test
*Feedback
*Research
9 – ASSEMBLING
(June-July 2017)
Administering the test (sept.2017)
1- SURVEY, PREPARATION specifications
syllabus research, creation, *kanji* list (Sep.2015)
2- DESIGN :
Task & Item writing (A1, A2, B1)
8 – Constructing ITEM BANK
3- Review
peer discussion, improvement, rejection, approval
7- STANDARD SETTING
7-10 expert panelists
(June 2017)
4 – VALIDATION 1st step : PILOTING, ANALYSIS classical testing theory, 100 test takers for A1 and A2 respectively, 50 for B1
«Think-aloud protocol»
6- VALIDATION 2nd step :
PRETESTING (Item response theory (IRT), Rasch Model)
500 test takers (250 pers. x 2 sets)
March –April 2017
5 – REVIEW : elimination of bias or rejection

**Figure 1**. The test development cycle of SELF (ALTE, 2011; Cervini, 2016)

The first step of the cycle is to research the available syllabuses for the creation of tests. Unlike European languages, Japanese does not have a completed CEFR-based syllabus. So we have had to develop our own syllabus referring to the CEFR descriptors and a few CEFR-based Japanese syllabi, and we have constituted lists of Chinese characters, kanji (see sections 3.2 and 4). Step 2 concerns the task and item writing, referring to CEFR descriptors and Japanese language characteristics. Step 3 is to review these tasks designed by test developers, to improve, approve or reject by peer discussion. Only approved tasks and items will be tested in step 4, “piloting”. During this step, we also collected data by think aloud protocol with two learners by level, which allowed us to conduct qualitative analysis. At Step 5, in the light of the result of statistical analysis (classical testing theory), the items with inadequate value were rejected. Step 6 is the last validation through pretesting with almost 500 learners of Level A1 to B1. The result is analyzed with Item response theory. Then we held standard-setting meetings to fix cut-off points and construct an item bank with validated items. The last step is to determine the algorithm and to assemble the test.

### 3.2 Authenticity of the tasks

Authenticity is the central notion of task conception in SELF (Cervini & Jouannaud, 2015). However, due to the Japanese graphic system, usage of authentic resources is difficult. It is important to know that the Japanese graphic system includes two systems of phonetic writing named hiragana and katakana, and another system of Chinese characters, named kanji (with the official kanji list containing about two thousands characters), and that these 3 writing systems are used conjointly even in a short sentence. Generally, about 50 kanji are taught at A1 level. So, for example, an A1 user can’t understand a simple notice at a railway station, because a lot of kanji are used for the proper name of the station, for example. The type and genre of text described as A1 or A 2 level, like posters, city maps, restaurant menus, is not usable without modification.

So, the majority of our tasks are fabricated or rewritten, respecting situational and interactional authenticity (ALTE, 2011). We focused on situational authenticity, that is “tasks and items representing language activities in real life” and created a similar text type or text genre that learners of Japanese in Europe should face in daily life. The majority of learners have never been to Japan, but they often practice online-based language activities such as social networking, blog chat, etc. (Project on Language Activities and Competences of the CEFR B1 level, 2012). With regard to interactional authenticity, that is “naturalness of the interaction between test taker and task and the mental processes which accompany it”, our tasks ensure the interaction (dialogue) is always between a native speaker and a non-native Japanese user. We also make sure that the test-takers can put themselves in the place of the non-native speaker. To create or rewrite the conversation, we referred to conversational analysis to ensure the naturalness of the scheme.

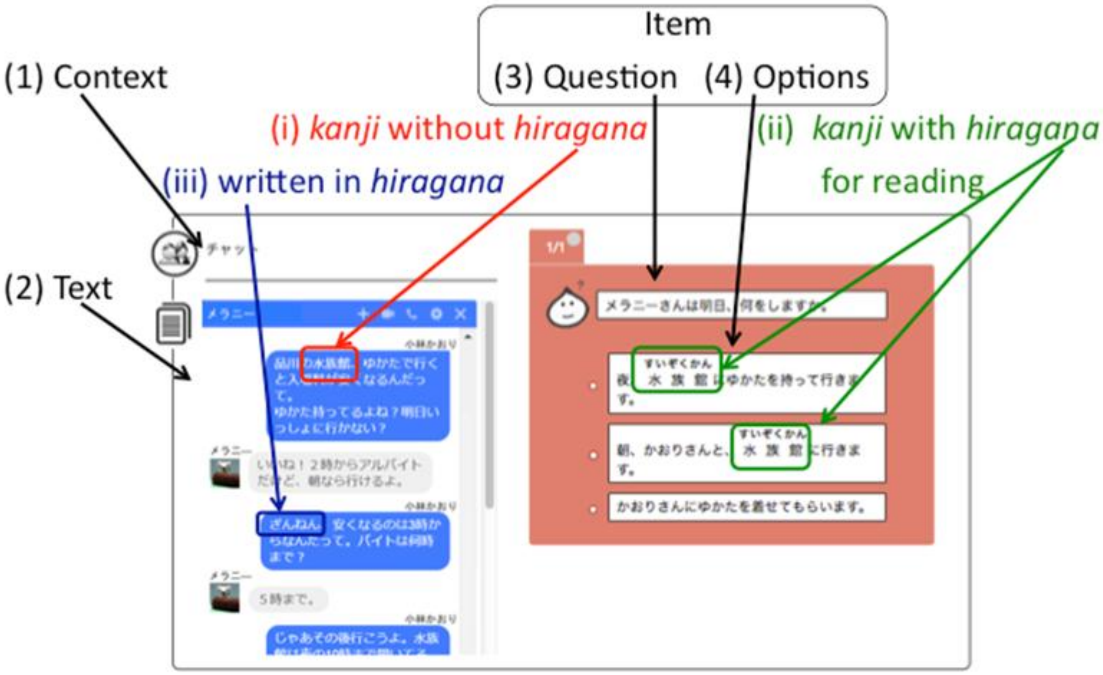
### 3.3 Reflection on sociocultural knowledge

From the viewpoint of the CEFR, the user/learner’s competences are sub-divided into two parts: “general competences” including declarative knowledge (such as knowledge of the world, sociocultural knowledge and intercultural awareness) (Council of Europe, 2001, pp. 110-112) on the one hand, and on the other, “communicative language competences”, which are further subdivided into three: Linguistics, Sociolinguistic and Pragmatic competences (ALTE, 2011.pp. 10-11). The SELF test assesses the communicative language competences of learners, but the task should not focus on declarative knowledge, which might distort the result on language competences. However, when we create tasks, with a view of authenticity and communicative approach, the sociocultural aspect is intrinsic in the text. Considering that Japan is culturally distant for European learners, some words or topics may cause sociocultural problems for the comprehension of text. We have chosen well known and ordinary words or notions such as manga, sushi, Kyoto, both stereotypic and explicit ones. But, sometimes, the learners don’t understand the underlying role and functioning of a word or a notion in Japanese society. In this case, the test doesn't assess learners' language competences. This problem is going to be discussed in the next section.

## 4 Task and item

Here are two examples of reading task related to cultural events in which people participate wearing a yukata, a kind of cotton kimono.

Figure 2 is the first example, a reading task of B1 level, "B1_CE_aquarium".



Item
(1) Context
(3) Question
(4) Options
(i) kanji without hiragana
(ii) kanji with hiragana for reading
(iii) written in hiragana
(2) Text
チャット
メラニー
小林かおり
1/1

品川の水族館、ゆかたで行くと入館料が安くなるんだって。
ゆかた持ってるよね？明日いっしょに行かない？

いいね！２時からアルバイトだけど、朝なら行けるよ。

ざんねん。安くなるのは3時からなんだって。バイトは何時まで？

5時まで。

じゃあその後行こうよ。水族館は夜の10時まで開いてる

メラニーさんは明日、何をしますか。

- 夜、水族館にゆかたを持って行きます。
- 朝、かおりさんと、水族館に行きます。
- かおりさんにゆかたを着せてもらいます。

**Figure 2**. Reading task B1, "B1_CE_aquarium"

The tasks in SELF are composed of four elements: (1) context, (2) text, (3) question, (4) options, and the last two elements make up an item. Some tasks have more than two items. For Japanese written tasks, we decided how to write the words in kanji (Chinese characters) in the task. We made a list of kanji for each level, 57 characters for A1, 144 for A2 and 211 for B1, mainly based on the frequency of kanji in our original tasks. Basically, we use only the kanji in the lists to write the words in the task (i), and for other words, we write them in kanji with small hiragana added above the kanji to help the test takers to read these words, called "furigana" (ii), or in hiragana (iii).

The translation in English of each part is as follows (the key words to answer the question are in square brackets):

(1) Context: chat

(2) Text

Kaori: Hi, at the [aquarium] at Shinagawa, the entrance fee will be discounted if we go there [in yukata].
You've a yukata, don't you? Shall we go [tomorrow]?
Mélanie: Great! I've a job from 2:00, but I can go in the morning.

Kaori: Too bad. It's [from 3:00] they discount. Until what time, your job?
Mélanie: [Up to 5:00].
Kaori: Well, let's go [after that], right? It's open until 10:00 [at night].
Mélanie: OK! You'll [help me to put on yukata]?
Kaori: OK!

(3) Question: What will Mélanie do tomorrow?

(4) Options: A. She will [bring a yukata] to the [aquarium at night].
B. She will [go to the aquarium] with Kaori [in the morning].
C. She will [have Kaori put on her yukata]. (key)

The text type is a chat between two friends, Kaori, a Japanese student, and Mélanie, a French student living in Japan. Kaori begins this chat to propose Mélanie to go to an aquarium with a special discount entrance fee for the visitors wearing a yukata. The key is the third option, C, which means that Mélanie will be helped by Kaori to put on her yukata tomorrow. The other key words to eliminate distractors are related to the time and the verbs.

This task is based on a descriptor of CEFR, B1 level for written interaction: "Can write personal letters giving news and expressing thoughts about abstract or cultural topics such as music, films" (Council of Europe, 2001, p .83). We set the context of this language activity, as a leisure activity concerning a cultural event in which they participate in yukata.

This concept is inspired by the fact that many events in yukata for international students in Japan are organized by their university or local association. In addition, as this event at the aquarium was really organized in Tokyo, we consider this context is situationally authentic. However, according to the survey data on yukata, 60 to 70 percent of young Japanese women cannot put it on by themselves. So, they usually ask someone to help them to get it on, using an expression in causative-benefactive form of the verb "put on". The context in which this expression of the function "asking for help" is used is so natural to us that we focused on it for this task.

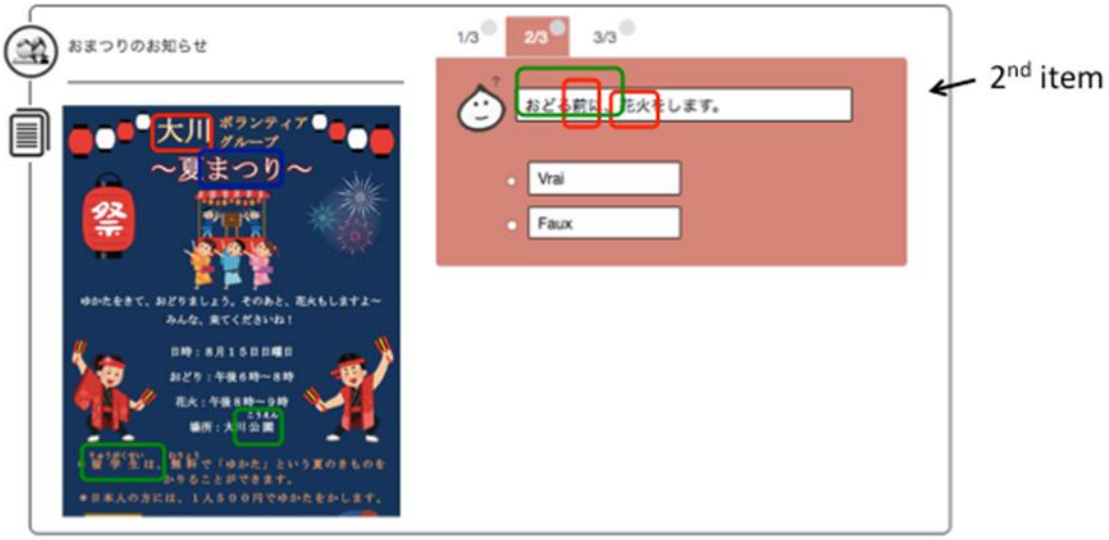
Contrary to our expectation, the results of the piloting test revealed that this item is too difficult for B1 level students, as the proportion of correct response is only 25 (see Table 1). We suppose that test takers believe that only children need to be helped to put on their clothes, at least in France where they do not wear a kimono or a yukata, except for a simple yukata as a nightdress when staying at a Japanese inn. This belief became a cultural bias that interfered with the test takers' comprehension.

The second example is a reading task of A2, "A2_CE_fete_d_ete", summer festival, of which the text is a festival poster (Figure 3). We also apply the same writing rules for words in kanji, written above, for the A level tasks. This task has 3 true-false type items, which the test takers can answer by clicking on the relevant numbers.

(i) *kanji* in the list (144) without *hiragana*: 大川, 夏, 来, 日時, 花火, etc.

(ii) *kanji* with *hiragana* for reading: 公園,留学生, 無料, etc.

(iii) written in *hiragana*: まつり（祭）, きもの（着物）, おどり（踊り）, etc.



おまつりのお知らせ

1/3 2/3 3/3

おどる前に花火をします。

- Vrai
- Faux

2nd item

大川 ボランティアグループ

～夏まつり～

祭

ゆかたをきて、おどりましょう。そのあと、花火もしますよ～

みんな、来てくださいね！

日時：８月１５日日曜日

おどり：午後６時～８時

花火：午後８時～９時

場所：大川公園

＊留学生は、無料で「ゆかた」という夏のきものをかりることができます。

＊日本人の方には、１人５００円でゆかたをかします。

**Figure 3**. Reading task A2, “A2_CE_fete_d_ete”

The translation in English of each part is as follows (the key words are in square brackets):

(1) Context: Information of summer festival

(2) Text  
Okawa Volunteer Group “Summer Festival”  
Let’s [dance wearing yukata]. [Then, do fireworks], too! Everyone, come to the festival!  
Time and Date: Sunday, August 15  
Dance: 6:00 p.m. – 8:00 p.m.  
Fireworks: 8:00 p.m. – 9:00 p.m.  
Place: Okawa Park  
*[The international students can borrow a “yukata”, a summer kimono for free].  
*We will [rent a yukata to the Japanese people at 500 yen].

(3) Question: the keys are in parenthesis.  
Item 1: You can [dance wearing summer kimono] in the festival. (True)  
Item 2: You will [do fireworks before dancing]. (False)  
Item 3: [The international students can borrow a yukata at 500 yen]. (False)

This task is designed based on a A2 level CEFR descriptor for reading activities: “Can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus, reference lists and timetables” (Council of Europe, 2001, p. 70). We imagined a festival poster organised by a local volunteer group as information of leisure activities concerning a cultural event in which they participate in yukata. As in the first example, this concept is inspired by the events in yukata organized in Japan, but to adapt to the language activities of A2 level learners, we made a bill for the international students. In this case, we can use some paraphrasing or additional explanations concerning traditional Japanese culture in the text just as in authentic posters in Japanese universities or associations to aid international students to understand Japanese culture.

The results of the classical item analysis of piloting data of these examples are shown in Table 1.

| | Difficulty (P-value) | Discrimination (Rir) | Options (A-value) |
|---|---|---|---|
| **Ex.1**<br>**B1_CE_aquarium** | 25 | 25 | 59 16 25* |
| **Ex. 2: item 1**<br>**A2_CE_fete_d_ete** | 80 | 31 | 80* 25 |
| **Ex. 2: item 2**<br>**A2_CE_fete_d_ete** | 80 | 49 | 20 80* |
| **Ex. 2: item 3**<br>**A2_CE_fete_d_ete** | 76 | 14 | 24 76* |

**Table 1.** Analysis of piloting data by Tiaplus (*key)/B1: 44 test takers, A2: 50 test takers (ALTE, 2011)

As we already mentioned, the proportion of correct response of the first example is 25, which indicates that this item is too difficult for B1 level, and the value of the discrimination index is inferior to 30. In contrast with the first example, the analysis data of the second example shows a high proportion of correct responses, 80 and 76, and the discrimination values of the first two items are superior to 30. These data mean that these items have a good validity as A2 level items.

Figures 4,5, 6 and 7 show the graphic data of two examples. The main factor that affects the discrimination of the first example is that the strongest test takers could not answer correctly, which may also be an evidence of a cultural bias.

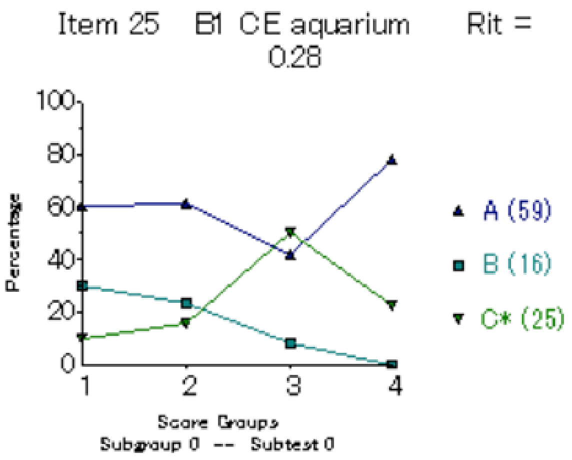


**Figure 4**. B1_CE_aquarium (*key)

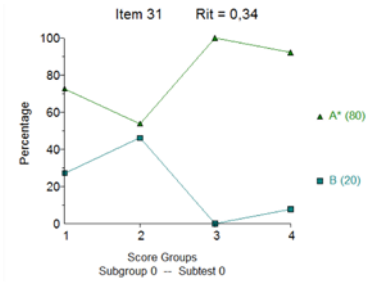


**Figure 5**. "A2_CE_fete_d_ete" item 1 (*key)

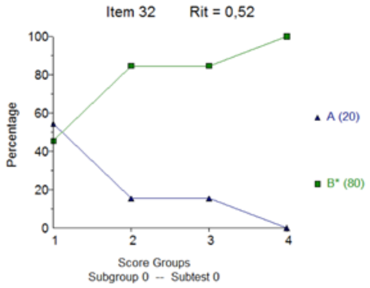


**Figure 6**. "A2_CE_fete_d_ete" item 2 (*key)

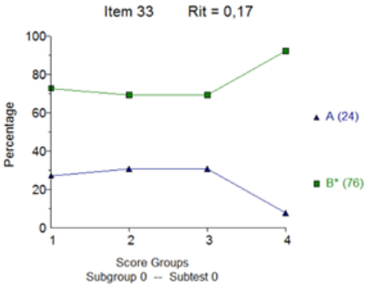


**Figure 7**. "A2_CE_fete_d_ete" item 3 (*key)

## 5 Conclusion

In this paper, we have demonstrated that the development of a Japanese test in a European environment and with the CEFR framework is possible but required some adjustments and consideration (Coste, 2007). SELF is a language test and assesses communicative language competence (with graphic, lexical, grammatical, sociolinguistic, pragmatic, and discourse components) but we showed that sociocultural knowledge plays a role in such communicative-type tasks. We have found that even an ordinary stereotypical sociocultural factor may distort an appropriate understanding of a situation if the learner didn't know the underlying functioning specific to the target culture. In other words, creating tasks with situational authenticity necessarily includes sociocultural factors. The higher the level, the more implicit and abstract the required sociocultural knowledge becomes. We therefore highlight the importance of "intercultural awareness", which would develop the sociocultural/intercultural competence, inseparable from communicative language competence (Byram, Zarate, & Neuner, 1997). This aspect should be taken into account in Japanese language education if it aims to use a real communicative/action -oriented approach.

## References

ALTE. (2011). *Manual for Language Test Development and Examining*. Strasbourg: Council of Europe.

Bachman, L.F. & Cohen, A. (1998). *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.

Byram, M., Zarate, G., & Neuner, G. (1977). *La compétence socioculturelle dans l'apprentissage et l'enseignement des langues*. Strasbourg: Council of Europe.

Cervini, C. (2016). Approcci integrati nel testing linguistico: esperienze di progettazione e validazione in prospettiva interlinguistica. In M.D. Miller (Ed.) *Interdisciplinarità e apprendimento linguistico nei nuovi contesti formativi. L'apprendente di lingue tra tradizione e innovazione* (pp. 64–85). Bologna: Quaderni del CESLiC.

Cervini, C., & Jouannaud, M. P. (2015). Ouvertures et tensions liées à la conception d'un système d'évaluation en langues, numérique, multilingue et en ligne, dans une perspective communicative et actionnelle. *Alsic, 18*(2). Retrieved from https://alsic.revues.org/2821

Coste, D. (2007). *Contextualiser les utilisations du cadre européen commun de référence pour les langues*. Paper presented at forum intergouvernemental sur les politiques linguistiques, Strasbourg, 6–8 February 2007.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

KU Leuven Faculty of Arts (2011) *Project on Language Activities and Competences of the CEFR B1 level (2012)*. Retrieved from http://japanologie.arts.kuleuven.be/en/research/project-language-activities-and-competences-cefr-b1-level

The Japan Foundation. (2003). *Survey report on Japanese Language Education Abroad 2003*. London: The Japan Foundation.

The Japan Foundation. (2017). *Survey report on Japanese Language Education Abroad 2015*. London: The Japan Foundation.

# How to Assess Mediation?

**Waldemar Martyniuk**, Jagiellonian University, Kraków, Poland

**Abstract:** This paper summarises student discussions on the newly developed illustrative descriptors for mediation, the fourth mode of language activities and strategies presented in the Common European Framework of Reference for Languages (CEFR) of the Council of Europe. The discussions were carried out during a seminar on Assessment in Language Teaching and Learning with students in a M.A. programme on Teaching Polish as a Foreign Language at the Jagiellonian University in Kraków, Poland, in the academic year 2016/17. The development of illustrative descriptor scales for mediation was carried out by a team lead by Brian North and Enrica Piccardo within the context of a wider project supported by the Language Policy Unit of the Council of Europe in Strasbourg.

## 1 Introduction

The main subject for discussions carried out during a seminar on Assessment in Language Teaching and Learning with students in a M.A. programme on Teaching Polish as a Foreign Language (henceforth PFL) at the Jagiellonian University in Kraków, Poland, in the academic year 2016/17, was the newly developed set of illustrative descriptors for mediation, the fourth mode of language activities and strategies presented in the Common European Framework of Reference for Languages (CEFR) of the Council of Europe. The descriptors were still in the developmental stage at that time and offered for piloting, so the main question guiding the discussions during the seminar meetings was to examine their applicability in the PFL area studied by the students. Reference Level B1 was selected for a more thorough investigation. Selected descriptors were translated into Polish and the students were invited to develop related teaching and learning activities as well as tasks for assessment. The observations from these group work activities were collected and reflected upon during plenary discussions – a summary of these is presented in the conclusions here.

## 2 The wider project

The development of illustrative descriptor scales for mediation was carried out by a team lead by Brian North and Enrica Piccardo within the context of a wider project supported by the Language Policy Unit of the Council of Europe in Strasbourg. The aim of the project was to provide an extended version of the CEFR scales based on validated CEFR-related descriptors that had been calibrated in a number of recent projects. In addition, a collation of CEFR-related descriptors for younger learners, mainly from European Language Portfolios, was prepared. The descriptor project was accompanied by a text by Daniel Coste and Marisa Cavalli aiming to “reposition the basic CEFR model within a more all-embracing view of social agents’ learning trajectory and personal development” (Council of Europe, 2015, p. 6).

All three descriptor projects mentioned above (2001 update; mediation descriptors, descriptors for young learners) have now (July 2017) been completed. The extended version of the CEFR illustrative descriptors, which integrates the results from the first two projects, was accessible in a preliminary, consultative edition on the Council of Europe (Language Policy Unit) website and will be officially launched in early 2018.

## 3 The seminar work

The work during the seminar was structured as follows:

- discussing the (updated) Descriptive Scheme of the CEFR and the rationale behind the development of the new scales for mediation
- translating the new mediation categories into Polish
- translating selected descriptors for B1 into Polish
- designing classroom activities in teaching PFL based on selected descriptors for B1
- designing suitable tasks and criteria for assessment.

Two tasks were to be carried out by the participants after the seminar:

- piloting selected classroom activities and assessment tasks in classes with B1 level students of PFL
- drafting recommendations for PFL area.

## 4 Initial discussions

The document presenting the newly developed and scaled descriptors for mediation (, 2016) that provides a brief introduction and (as appendices) rationales for each of the scales, proved to be extremely inspiring for the introductory discussions on the CEFR approach and its Descriptive Scheme. Two major striking observations were noted by the participants here: the overarching importance of the so far undervalued or even neglected mode of mediation integrating all other language activities and strategies, and the applicability of a majority of the mediation scales to language education beyond the area of foreign languages, reaching out to the teaching and learning of any school subject. The first observation was evoked by the diagram reproduced by North & Piccardo as Figure 1 that appeared in the 1998 version of the CEFR. In this schema, mediation is shown as an extension of interaction, which in turn connects aspects of reception and production. This figure highlights the high value of mediation, clearly indicating how it integrates the other language activities and strategies reflecting the complex nature of each act of communication.

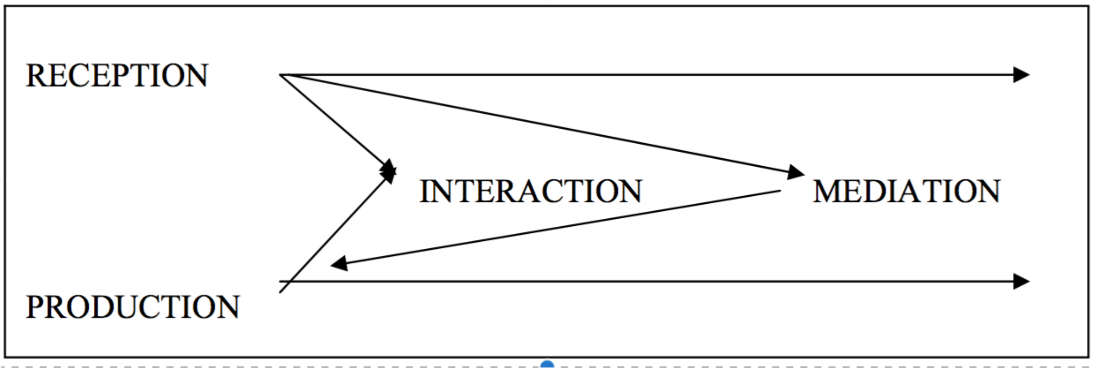


**Figure 1.** Mediation in the CEFR, 1998 edition (North  & Piccardo, 2016, p. 4)

The second observation was based on both the document edited by North and Piccardo and the Coste and Cavalli text on the mediation functions of schools. North and Piccardo underline that their interpretation of mediation is "more in line with educational literature within and beyond the language field (which) leads to a definition of mediation competences that are potentially relevant to all types and contexts of language use” (North & Piccardo, 2016, p. 3). They admit that it is “a significant, and deliberate, departure from the targeting of the original illustrative descriptors, which were specifically designed in relation to the foreign/second language classroom only” (North & Piccardo, 2016, p. 3). Their interpretation is fully in line with the claim made by Coste & Cavalli that “although the CEFR was designed, and has been used, above all in relation to the learning of foreign languages, it presents a model that is just as valid for all other forms of language communication” (Council of Europe, 2015, p. 10).

## 5 Mediation in the CEFR 2001

Mediation is presented in the CEFR original version published in 2001 as the fourth group of activities carried out and strategies applied by a language user/learner after reception (listening and reading), speaking (interaction and production), and writing (interaction and production) defined in the following way (Council of Europe, 2001, p. 14):

> In both the receptive and productive modes, the written and/or oral activities of mediation make communication possible between persons who are unable, for whatever reason, to communicate with each other directly. Translation or interpretation, a paraphrase, summary or record, provides for a third party a (re)formulation of a source text to which this third party does not have direct access. Mediating language activities – (re)processing an existing text – occupy an important place in the normal linguistic functioning of our societies.

No reference scales of descriptors were provided in the text to illustrate mediation activities and strategies with only some examples of related language use given, such as:

- oral mediation: simultaneous, consecutive, informal interpretation
- written mediation: exact and literary translation, summarising, paraphrasing

North and Piccardo summarise the concept of mediation as presented in the CEFR 2001 as follows (Council of Europe, 2015, p. 6):

In mediation activities, the user/learner:

- receives a text and produces a related text to be received by another person who has no access to the first text
- acts as an intermediary in a face-to-face interaction between two interlocutors who do not understand one another, possibly because they do not share the same language or code
- interprets a cultural phenomenon in relation to another culture
- participates in a conversation or discussion that involves several languages, exploiting his/her plurilingual and pluricultural repertoires.

The seminar participants noted with interest that whereas the first two sentences in the original CEFR 2001 mediation definition did indeed refer to actions to be performed between users of two different languages, the final sentence about incorporating and (re)processing an existing text in the definition opens up a vast territory for mediation to encompass actions undertaken between any two (or more) texts, in the same or in two (or more) different languages. In this context, an insider remark made (most probably) by North is worth quoting (Council of Europe, 2015, p. 7):

> Looking at what has happened from a historical perspective can also help to cast light onto this development. Two main considerations in particular seem to be helpful. The first is the fact that North's (1992) category 'processing' in the presentation of the schema reception, interaction, production, processing at the 1991 Rüschlikon Symposium that recommended the development of the CEFR and ELP, was replaced by the category 'mediation' during the work of the CEFR's Authoring Group. It is possible that the authors continued to be over-influenced by this association with processing text. Descriptors for processing were in fact developed during the Swiss National Research Project referred to above – but then included in Section 4.6.3 'Text,' rather than under mediation.

Mediating/acting between language users as well as between texts were the two aspects of mediation explored in full first with the development of the new scales of illustrative descriptors.

## 6 Mediation in CEFR 2017 (consultative version)

The scope of mediating actions opened up by the original CEFR 2001 definition was significantly extended by Coste & Cavalli in their text on mediation functions of schools. Their definition of mediation highlights its essential nature as a communicative language activity that applies to a very wide range of contexts (Council of Europe, 2015, pp. 62-63):

> To mediate is, inter alia, to reformulate, to transcode, to alter linguistically and/or semiotically by rephrasing in the same language, by alternating languages, by switching from oral to written expression or vice versa, by changing genres, by combining text and other modes of representation, or by relying on the resources – both human and technical – present in the immediate environment. Mediation uses all available means and this is its attraction for language learning and the development of a range of discourse competences.

A comment made here by several seminar participants was that taking all of this into account leads to a conclusion that, in principle, mediation is at the core of every educational act, every instance of learning. Indeed, North & Piccardo are perfectly aware of the fact that their mediation descriptors may have much broader application beyond teaching and learning of foreign languages targeted by the original CEFR descriptors (North & Piccardo, 2016, p. 44):

> This is breaking new ground. Unlike with the original CEFR illustrative descriptors, or the other two related descriptor projects (updating the 2001 scales; collating descriptors for young learners) the focus (here) was not on foreign languages. (…) Nowadays, given the level of mobility and migration and the variety of ethnicities in city classrooms, the notion of native speaker and even the dichotomy language of schooling/foreign language, let alone mother tongue/foreign language, loses its validity. This is one of the main reasons that expressions like mother tongue, second language, source language, target language, etc. are not used in the mediation descriptors. It is simply suggested that the user should name the precise languages involved.

The set of scales of illustrative descriptors developed and validated for the above outlined concept of mediation and offered for consultation in 2015-17 included the following categories:

- Relational mediation
- Cognitive mediation
    - Constructing meaning (spoken)
    - Conveying received meaning (spoken)
    - Conveying received meaning (written)
- Mediation strategies

The initial set used for piloting included a total of 24 scales reduced after the first round of consultations to 19 – still quite a considerable extension of the original 2001 set of 54 scales covering *Reception*, *Production*, and *Interaction*. The three major groups of scales – *Relational mediation* (acting between language users), *Cognitive mediation* (working with texts), and *Mediation strategies* – were subdivided into the following sets of can-do statements (the ones indicated in italics were dropped after the first round of piloting):

Relational mediation

- *Establishing a positive atmosphere*
- Creating pluricultural space
- *Facilitating collaborative interaction with peers*
- Managing interaction in plenary and in groups
- Resolving delicate situations and disputes

Cognitive mediation

- Constructing meaning (spoken)
    - *Collaborating to construct meaning*
    - *Generating conceptual talk*
    - Stimulating development of ideas
- Conveying received meaning (spoken)
    - Relaying specific information
    - Explaining data (in graphs, charts etc.)
    - Processing text
    - Interpreting
    - Spoken translation of written text (Sight translation)
- Conveying received meaning (written)
    - Relaying specific information
    - Explaining data (in graphs, charts etc.)
    - Processing text
    - Translating

Mediation strategies

- Linking to previous knowledge
- Amplifying text
- Streamlining text
- Breaking down complicated information
- *Visually representing information*
- Adjusting language

## 7 Application of the new descriptors for PFL

The seminar participants decided to select the reference Level B1 for closer investigation to examine the applicability of the new descriptors in the PFL area. They noted immediately that the two aspects of mediation – acting between people and working with texts – are perfectly well reflected in the overall scale illustrating the ability at this level, with two descriptors representing each of the two facets. Translating the categories and the selected B1 can-do statements into Polish offered the seminar participants an excellent opportunity to reflect on the possible operationalisation of the individual descriptors in terms of activities for teaching and learning of PFL, as well as in terms of tasks for assessment. While the descriptors related to working with texts (*cognitive mediation* and *mediation strategies*) were easier to tackle, the ones related to actions to be performed between people (*relational mediation*) seemed to be far more difficult to turn into classroom activities and tasks for assessment. The following sets of descriptors were perceived as the most problematic ones in this respect, specifically in terms of creating tasks and criteria for assessment:

- Relational mediation, creating pluricultural space
- Relational mediation, managing plenary and group interaction
- Relational mediation, dealing with delicate situations and disagreements.

## 8 Conclusions – a summary of issues

The observations noted during the seminar activities were collected and discussed during plenary discussions. The following summary reflects issues raised during these discussions and may serve as a conclusion to this report:

- Can mediation as language activity really be extended beyond text processing/transforming?
- Is mediation between languages (plurilingual mediation) really the same as within one language (L1 or L2)?
- Is relational mediation ability building strictly on communicative language competences or rather, more directly, dependent on general competences (savoirs, specifically the existential competence, savoir-être)?

- Is this newly conceptualised mediation to be still perceived as a skill defined in terms of activities and strategies or rather a competence in its own right required to be defined more broadly (mediative language competence)?
- Are not reading and listening comprehension possible to assess through mediation activities/tasks only?
- Are not relational and cognitive mediation activities and strategies in fact defining what education in general is all about?
- How can relational mediation be assessed?

## Further reading

Language Policy Unit. (2016). *Piloting new descriptor scales from a proposed extended version of the CEFR illustrative descriptors.* Retrieved from https://rm.coe.int/1680703acd

## References

Council of Europe. (1998). Modern languages: Learning, teaching, assessment: A Common European Framework of reference, Council for Cultural Co-operation, Education committee, CC-LANG (95)5 rev. V.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2015). *Education, mobility, otherness. The mediation functions of schools*. Retrieved from http://www.coe.int/t/dg4/Linguistic/Source/LE_texts_Source/LE%202015/Education-Mobility-Otherness_en.pdf

North, B. & Piccardo, E. (2016). *Developing illustrative descriptors of aspects of mediation for the Common European Framework of Reference (CEFR)*. Retrieved from rm.coe.int/developing-illustrative-descriptors-of-aspects-of-mediation-for-the-co/1680713e2c

# Language learning, teaching and assessment…

# … in the digital era

# The Impact of the Integrated Teaching, Learning and Assessment Framework on Students’ Writing Perceptions and Performance

**Huang Jing**, China West Normal University, China

**Abstract:** In order to address the conflict between a process writing approach to teaching and a summative assessment strategy, this research proposed a four-stage integrated teaching, learning and assessment framework and then explored its influence on students’ perceptions and performance. Multiple sources of data were collected in a natural learning setting for one semester. Through the analysis of students’ perceptions of writing learning, results show that due to more interactions, their anxiety caused by writing and revising relaxed, their writing confidence was enhanced and they self-reported writing progress; Through the analysis of students’ writing performance, the participating students wrote more drafts, paid more attention to language use after receiving multiple-sourced feedback, revised more by themselves, and improved the quality of their writing products to different degrees. The framework succeeded in solving the conflict by creating a feedback-rich environment for individualised autonomous learning.

## 1 Introduction

In order to address the conflict between a process-writing approach to teaching and a summative assessment approach on writing in a tertiary EFL learning context in Mainland China, this research proposed a 4-stage integrated teaching, learning and assessment framework and explored its influence on students’ learning to write in two dimensions: perceptions and performance.

## 2 A problem identified between teaching and assessment

In Chinese College English writing classrooms, writing as a language skill has been mostly carried out in an examination-oriented approach that emphasises summative assessment. In the traditional writing classroom (see Lo & Hyland, 2007), after assigning a writing topic, the teacher provides some input on the grammatical structures and vocabulary needed for the writing task. Students then write within a time limit with a certain word limit, very much with the teacher in mind as the audience, and submit single drafts for teacher comments. In the traditional summative feedback approach to writing, teachers and students are engaged in an unequal relationship, with the teacher having absolute authority over students’ writing products and students only passively obeying the teachers’ directions and requirements for changes in their texts without knowing what kind of criteria or standards the teachers used to judge their writing.

## 3 The integrated teaching, learning and assessment framework

The framework described in this section synthesized two different bodies of literature: the Assessment for Learning literature and social learning theories. As the result, the integrated framework is intended to realise what Lee & Coniam (2013) pointed out: “to implement Assessment for Learning in writing classrooms effectively, two pedagogical practices are conductive to student learning: firstly, helping students understand assessment criteria; and secondly providing feedback and encouraging reflection on it” (p. 46). This emphasises the active role the students play in their own learning in the assessment and feedback process, and writing and rewriting process supported by different feedback providers and effective instructional support, and through self-reflection centered on the assessment criteria. The influence of social learning theories of language development has focused attention on the interactive and collaborative aspects of feedback and its crucial dialogic role in scaffolding learning (Hyland, 2010).Thus, this study presupposes that the effectiveness of feedback depends on the learning theory mainly adhered to in social learning theories, such as Zone of Proximal Development, Community of practice, and Meme theory.

The framework presupposes the need to refocus feedback research and practice away from the notion of instructors providing one-way feedback to students in favor of dialogic exchanges in which instructors and students are jointly involved in conversations about learning. Central to the rationale is a desire to provide a mechanism through which the students could acquire and utilise evaluation skills to improve their learning. In contrast, it is a system where the evaluation of the process complements that of the product, where learning and assessment concur, and where learner independence is openly fostered. Unless students are developing capacities to self-regulate their own learning, their ability to make sense of and use any feedback provided is seriously constrained. Therefore, from the outset, the integrated framework is seen as having three main aims:

(1) To resolve the incoherence between a process writing approach to the teaching and a summative assessment approach through a one-shot feedback procedure.

(2) To assess students' writing process as well as the product of the learners' writing with a multiple-draft process-oriented approach to writing and the complexity of the many aspects involved in the dialogic communication between writing teachers and student writers.

(3) To facilitate learners' independence by improving their awareness of both their writing techniques and the standard criteria of the specific genre.

Since writing is a dynamic and multi-stage process, effective assessment and feedback procedures should be dynamic and multi-staged accordingly to facilitate students' writing development, which means assessment and feedback on student writing should be tangible and supplied to students at different stages of their rewriting and revising processes.

Thus, following the literature review and the recognition of the points summarized above, the integrated framework is formulated in four distinct stages: Stage 1 is the activating lecture to clarify what good performance is by introducing the task requirement and its assessment criteria to prepare students to write; Stage 2 is the integrated feedback process to influence student writing by identifying the gap between expected performance and students' current performance through the use of feedback such as self, peer, teacher and computer-generated feedback in combination on one task; Stage 3 is the teaching and assessment workshop to reinforce the assessment criteria, and further emphasise the strengths and weaknesses of student drafts through instructor's overall report, peer review presentation by a group and sample analysis; and Stage 4 is self-assessment, to help students to internalise the assessment criteria for the expected performance of specific tasks with reference to their own writing performance.

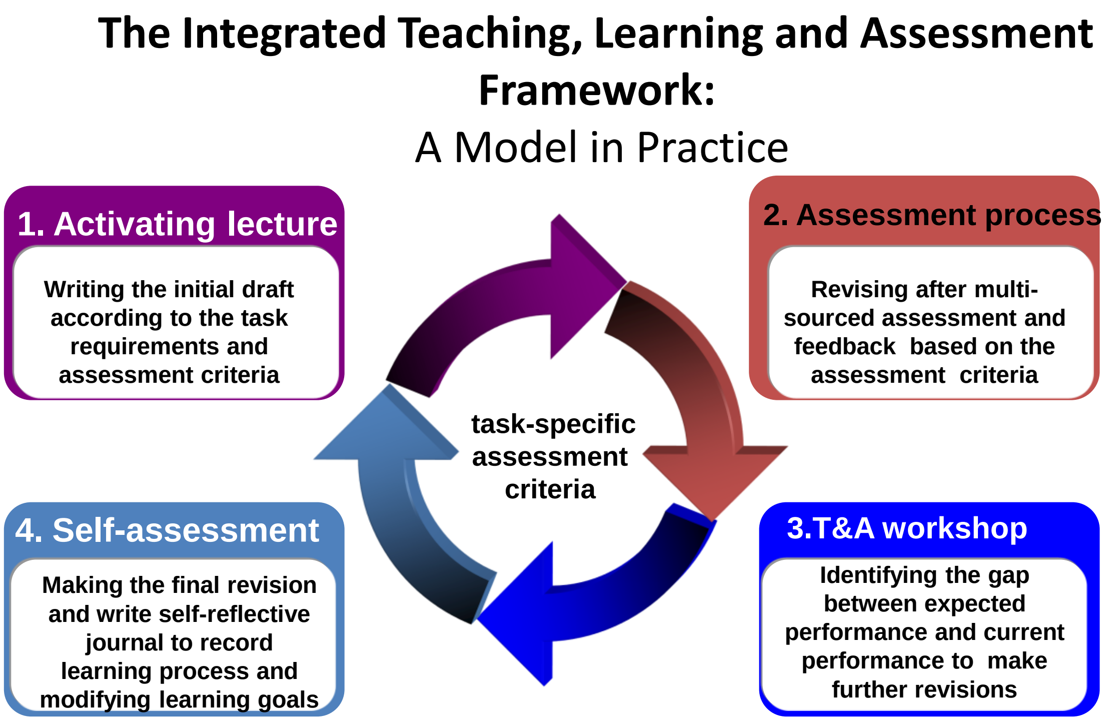

The Integrated Teaching, Learning and Assessment Framework:
A Model in Practice
1. Activating lecture
Writing the initial draft according to the task requirements and assessment criteria
2. Assessment process
Revising after multi-sourced assessment and feedback based on the assessment criteria
task-specific assessment criteria
3.T&A workshop
Identifying the gap between expected performance and current performance to make further revisions
4. Self-assessment
Making the final revision and write self-reflective journal to record learning process and modifying learning goals

**Figure 1**. The integrated teaching, learning and assessment framework

### 3.1 Stage 1: the activating lecture

In the activating lecture, what good performance is with reference to the assessment criteria is revealed to the students. The assessment criteria is one of the cornerstones of the integrated framework which allows learners to compare their work against standards, and has the advantage of encouraging learner independence with support from the integrated feedback process in the second stage. If the task requirement and criteria are not well specified, it is difficult for teachers and students to understand students' accurate ability level and to expand students' zone of proximal development (Vygotsky, 1978).

### 3.2 Stage 2: the integrated feedback process

The integrated feedback process is to help students to identify the gap between the expected performance and their current performance by making revision an organic part of feedback. In other words, redrafting, self and peer evaluation have to be built into the writing process, and students should be prepared to deal with subsequent drafts themselves. Educational research suggests that feedback is more effective when information is gathered from the subjects themselves as well as others (Brinko, 1993). During this process, feedback is given to the students' writings from human sources (peer, teacher, self) with reference to the task-specific assessment criteria related to the goals of the course instruction and task specification, and as well as from computer-generated feedback. Feedback includes not only micro-level corrective feedback addressing spelling, grammar, word choice, and missing words, but also macro-level comments that address paper organisation, quality of the ideas contained, and other larger levels. Dealing with these larger idea-and-argument-centered comments may encourage students to improve the quality of the larger issues in writing and prevent them from focusing on the smaller technical issues of writing.

### 3.3 Stage 3: the teaching and assessment workshop

As an extension of the activating lecture (the first stage) and the integrated feedback process (the second stage), the teaching and assessment workshop could help students understand the task-specific assessment criteria in depth after experiencing the integrated feedback engagement process, and the strengths and weaknesses of student drafts with instructional supportive strategies related to different feedback activities such as teacher's evaluative report of the whole class, one or two peer review group presentations and a analysis of a few sample pieces.

### 3.4 Stage 4: self-assessment

As the last phase, self-assessment is a process of wrapping up the whole learning process for one task through self-reflection. With the support of the task-specific assessment criteria, the self-evaluation through reflection process closes all the links of the integrated framework. A key feature of the framework that differentiates it from others is that students are assumed to occupy a central and active role in all the assessment and feedback related activities. Students receive computer-generated feedback, peer feedback and teacher feedback with reference to the assessment criteria in order to further understand and identify their gap between the expected performance and their current performance prior to self-evaluation through reflection. They are virtually directed to complete the cycle of "learning" by understanding and engaging with assessment and feedback from different processes to inform their performance self-evaluation.

## 4 The research design and findings

The central concern in this research is to explore how the integrated framework influences Chinese tertiary EFL writers' learning how to write. More specifically, it focuses on how it influences EFL writers' perceptions and performance. In an attempt to address this research issue, three research questions are addressed in this paper:

(1) To what extent and how did the integrated framework have an effect on students' perceptions of learning to write in English?

(2) To what extent and how did the integrated framework have an effect on students' writing performance?

(3) What were the factors that influenced the implementation of the integrated framework ?

A mixed methods research design was applied longitudinally. 25 students were involved in the research over one semester in a natural classroom setting. The study was conducted with a triangulation of data collection and analysis. The multiple sources of data included pre-and post-course writing tests; student writing samples; feedback from different feedback processes including computer-generated feedback, peer feedback, teacher feedback and self-feedback; text revision; reflective journals; student questionnaires; student and instructor interviews; and researchers' classroom observation.

In order to match the integrated teaching, learning and assessment framework, the integrated feedback procedures were designed. In essence, there were five steps to proceduralise the integrated feedback carried out in the course. The whole process is illustrated in Table 1.

### 4.1 Step 1: computer-generated feedback

After completion of their writing, students would first submit their initial draft to the Pigai system (www.Pigai.org) for the spontaneous diagnostic feedback, mainly focusing on linguistic forms. They worked with the program independently to revise their writing according to the automated feedback they received for each draft. The instructor did not require them to achieve a minimum satisfactory score or limit the number of submission times before essays were submitted for human feedback.

### 4.2 Step 2: Peer feedback

After revision based on the computer-generated feedback, each student would send his or her revised draft to his or her two peers for feedback with the guidance of the task-specific peer review form in accordance with the task-specific assessment criteria. Students were then required to incorporate peer feedback into subsequent drafts and a deadline was set for them to submit their essays for teacher feedback; meanwhile they needed to prepare a peer review group presentation for the whole class.

### 4.3 Step 3: Teacher feedback

After revision based on peer feedback and discussion, one piece out of each group were recommended and sent to the instructor, who evaluated one-third of the whole class in different modes such as one-to-one written feedback and 1-minute recorded oral feedback in "The Tripartite Evaluation Model". Before students had the second class meeting with the instructor, they were required to have read and listened to teacher feedback in both modes.

### 4.4 Step 4: Feedback workshop

In the feedback workshop during the second class meeting, with reference to the assessment criteria of the specific task, students were allowed to have a general picture of the performance of classmates by listening to the instructor's evaluative report of the writing performance of the whole class, listening to one or two peer review group presentations, and reviewing several sample pieces from fellow students.

### 4.5 Step Five: Self-reflection

After the second classroom meeting, students were invited to finalise their articles by synthesizing any feedback they received to correct language errors, and improve coherence, cohesion, relevance of ideas, and so on. Once the above mentioned procedures have been completed, students were required to complete a task-specific reflective journal, according to the course syllabus and the outline provided by the researcher. Students then submitted the final version and the reflective journal together to the teacher for learning process.

<table>
<tr><th>N</th><th colspan="3">Feedback sources</th><th>Criteria</th><th>Feedback modes</th><th>Interaction</th></tr>
<tr><td>1</td><td colspan="3">The computer-generated feedback</td><td>The pre-set standard in the databases</td><td>Written</td><td>Student-computer</td></tr>
<tr><td>2</td><td colspan="3">Peer feedback+ peer presentation</td><td>Course criteria<br>Task-specific criteria</td><td>Written</td><td>Student-student</td></tr>
<tr><td rowspan="2">3</td><td rowspan="2">Teacher feedback</td><td colspan="2">Out of class one-to-one written and recorded oral feedback</td><td rowspan="2">Course criteria&amp;<br>Task-specific criteria</td><td rowspan="2">Written/<br>Oral recorded</td><td rowspan="2">Teacher-student</td></tr>
<tr><td colspan="2"></td></tr>
<tr><td rowspan="3">4</td><td rowspan="3">Classroom Support (Teacher)</td><td rowspan="3">In class oral feedback to the whole class</td><td>Evaluative report for the whole class</td><td rowspan="3">Course criteria<br>Task-specific criteria</td><td rowspan="3">Written and oral</td><td rowspan="3">Teacher-class<br>Teacher-group<br>Teacher-student</td></tr>
<tr><td>Peer review presentation</td></tr>
<tr><td>Sample pieces</td></tr>
<tr><td>5</td><td colspan="6">Self-reflection +self-review + self-revision</td></tr>
</table>

**Table 1**. The integrated feedback procedures in “Happy English Writing”

Findings from this exploratory study suggest that the integrated framework has the potential to help learners to change their perceptions of writing learning and improve their writing performance. In many ways, this study confirms the findings of past research and provides additional insights into how teaching, learning, and assessment should be integrated in any writing classroom. Through the analysis of students’ perceptions of writing learning, results show that due to the guidance of the task-specific assessment criteria and more interactions between the teacher and student, student and student, teacher and text, student and text, self and text, and student and machine, their attitudes towards writing and revising changed, and their self-reported writing progress and writing confidence was enhanced. Through the analysis of students’ writing performance, the participating students wrote more drafts, revised more by themselves, paid more attention to language use after receiving multiple-sourced feedback, and improved the quality of their writing products to different degrees. By investigating students' perceptions and performance, we can identify a number of factors (proficiency levels, group dynamics, etc.) that may affect the implementation of the integrated feedback approach and contribute to the dynamics of classroom learning teaching, and assessing.

Limitations have been identified in the research design in two aspects. The first one is that there was not a control group. However, the variables of the control group are very hard to control, thus, the idea of using a control group was rejected. Moreover, it is a study with an exploratory nature and it does have enough data to address the research focus. The second problem is that students' perceptions should be measured by a pre-course questionnaire to form a comparison with the end-of-course questionnaire. In addition, the views only represented those of a sample of Chinese students in this particular research context. The extent to which they may represent students in other contexts is debatable.

## 5 Conclusions

Despite these limitations mentioned above, this study contributes to the literature in proposing a four-stage integrated teaching, learning and assessment framework and explored its impact on students' autonomous writing learning. To conclude, the integrated framework has a positive impact on students' perceptions of learning how to write as well as their writing performance; it promotes a community learning classroom culture as well as strengthens students' confidence as L2 writers. It succeeds in creating a feedback-rich environment for individualized autonomous learning, without total dependence on teachers. Therefore, the inconsistency between the teaching of process writing skills and a summative assessment approach through a one-shot feedback procedure was resolved.

While the current study has attempted to answer the research questions posed earlier, it has also brought out issues and research directions that need to be further explored in the future. In the future, teacher assessment and feedback may still have an important part to play, but placing too much credence in its powers of influence can lead to students' overdependence on teachers' assessment and feedback. As a result of attempts to overcome the limitations of teacher assessment and feedback, internationally, there is an emerging recognition of benefits associated with the use of alternative methodologies that shift a proportion of the responsibility for assessment to the students in higher education. A far-reaching conclusion, and one that the research has gradually drawn as it has proceeded, is that rather than focusing so much attention on providing assessment and feedback and seeing this as a central part of our identities as writing instructors, with help of the educational technology, we should perhaps be devoting more attention to developing students' assessment literacy to provide assessment and feedback among themselves and to themselves.

## References

Brinko, T. (1993).The practice of giving feedback to improve teaching, what is effective? *The Journal of Higher Education, 64*(5),574–593.

Hyland, F. (2010). Future directions in feedback on second language writing: overview and research agenda. *International Journal of English Studies*, 10(2), 171–182.

Lee, I. & Coniam, D. (2013).Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. *Journal of Second Language Writing, 22*(1), 34–50.

Lo, J. & Hyland, F. (2007). Assessment for learning: Integrating assessment, teaching, and learning in the ESL/EFL writing classroom. *The Canadian Modern Language Review, 64*(1),199–213.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.

# Análisis de la Severidad de los Calificadores de una Prueba de Expresión Escrita en el Contexto de una Prueba de Dominio

**Juan Miguel Prieto Hernández**, Cursos Internacionales, Universidad de Salamanca, España

**Abstract:** Esta comunicación describe cómo analizar el comportamiento de los calificadores de la prueba de expresión e interacción escritas de una prueba de dominio. Para ello, proponemos utilizar uno de los modelos de Rasch, el denominado Many-Facet Rasch Measurement (MFRM), con el fin de situar en la misma métrica los elementos de las facetas incluidas en el contexto de una evaluación de desempeño: candidatos, calificadores, tareas y atributos evaluados.

Para poder utilizar este método es preciso que cada examen sea calificado por dos examinadores que actúen de manera independiente, y también requiere crear una red de calificadores en la que queden conectados pruebas, candidatos y examinadores, de modo que se puedan comparar todos los datos en un mismo marco de referencia. De este modo, es posible lograr un escalamiento conjunto según el nivel de competencia de los candidatos, conocer la severidad/benignidad de los calificadores y determinar la dificultad de las tareas y de los atributos evaluados. El procedimiento también permite analizar la existencia de tendencia central o de efecto de halo en las calificaciones.

## 1 Introducción

Uno de los aspectos críticos de los tests de desempeño o de ejecución (*performance assessment*) (Martínez, 2010, p. 85) es la correcta asignación de las puntuaciones a las tareas cumplimentadas por los candidatos. Las puntuaciones que obtienen los candidatos que han realizado una prueba de desempeño no dependen únicamente del nivel de los examinados en el constructo de interés. La utilización de las escalas de calificación se basa en la suposición de que un examinador es capaz de realizar una correcta observación cuantitativa con precisión y con un cierto grado de objetividad (Guilford, 1954, p. 278). Las tres posibles fuentes de error que pueden poner en peligro la calidad de un proceso de calificación son: la utilización de las escalas de calificación, el proceso de calificación y el comportamiento de los examinadores (Popham, 1990). El objetivo de nuestra presentación fue centrarnos en el estudio del comportamiento de los examinadores en el proceso de calificación de una prueba de expresión e interacción escritas (EIE) del nivel A2. En concreto, nos hemos centrado en los errores relacionados con la severidad o benignidad de los examinadores (Myford y Wolfe, 2004, p. 471). Este tipo de error, según Cronbach (1990), es el más grave que un evaluador puede introducir en un procedimiento de calificación ya que las puntuaciones de los examinadores siempre deberían estar relacionadas con el nivel de competencia de los candidatos.

Las causas que pueden influir en que un examinador valore con mayor o menor severidad la actuación de un candidato pueden ser diversas: la cantidad de tareas que tenga que calificar y el tiempo del que disponga para hacerlo, factores idiosincrásicos como la personalidad del calificador y su actitud ante el proceso de calificación en el que va a participar, su experiencia, etc. (Eckes, 2011, p. 55).

## 2 Método

Los análisis y los datos expuestos durante la presentación forman parte de un proyecto de investigación realizado en la Universidad de Salamanca en colaboración con el Instituto Cervantes (Prieto, J.M., 2016). Para dicho estudio, se utilizaron datos de las puntuaciones que un equipo de doce examinadores otorgó, de forma individual, a los candidatos que se presentaron al examen para la obtención del Diploma de Español Nivel A2 en la convocatoria de mayo de 2012. De los 4301 candidatos que se presentaron a la prueba, el 88,7% (n=3858) fueron calificados siguiendo el método "tradicional" de reparto de exámenes entre los calificadores: se formaron parejas de calificadores y a cada integrante se le asignaron los mismos exámenes; uno calificó los exámenes originales y el otro fotocopias de los mismos. Para la calificación del 10,3% restante (n=443) se utilizó un procedimiento diferente que se detallará más adelante.

## 3 Instrumento

La prueba de EIE del Diploma de Español Nivel A2 se presenta en un único cuadernillo en el que aparecen las tareas a partir de las cuales se deben redactar las respuestas. La prueba consta de tres tareas: dos de interacción y una de expresión. La duración total es de 50 minutos. La extensión total de palabras que los candidatos deben escribir en el espacio reservado para cada tarea entre los tres textos oscila entre 170 y 200 palabras.

## 4 Procedimiento

Los examinadores puntuaron el rendimiento en cada atributo en una escala de 0 a 3. A cada una de las categorías (0, 1, 2 y 3) corresponde un único descriptor ilustrativo que se compara con la actuación del candidato. Se evalúan tres atributos, uno holístico y dos analíticos: (1) adecuación al género discursivo; (2) coherencia; y (3) corrección y alcance. En la calificación analítica de la prueba de expresión e interacción escritas, las tres tareas se ponderan de la siguiente manera: tarea 1 (17%), tarea 2 (33%) y tarea 3 (50%).

En los tres atributos, la categoría de 2 puntos es equivalente a la descripción del nivel A2 (*plataforma*) del *Marco común eropeo de referencia para las lenguas* (MCER), el valor 3 supone una consecución por encima del nivel, la categoría 1 supone la no consecución del nivel, y el valor 0 supone que la prueba está en blanco, que no sigue los puntos de orientación dados, que el candidato escribe información irrelevante que no se ajusta al objetivo planteado o que el texto es ilegible.

## 5 Modelos de análisis de las evaluaciones mediadas por calificadores

En la comparación de los promedios de calificaciones de los examinadores que calificaron el 88,7% de los candidatos (siguiendo el método "tradicional" de reparto de exámenes entre los calificadores), se constató que la ausencia de varianza de las propiedades de los tests respecto de los sujetos utilizados para estimarlos significaba que las puntuaciones otorgadas por los calificadores estaban relacionadas con el nivel de competencia de los candidatos. Como consecuencia, únicamente se encontraban en la misma escala los examinadores que habían calificado el mismo bloque de pruebas (originales y fotocopias) y no era posible realizar comparación alguna con el resto de calificadores. Diferencias significativas en las medias de puntuación parciales y globales entre calificadores podrían ser debidas a que habían calificado a candidatos con distintos niveles de competencia y no a que fueran más severos o benévolos.

Estas limitaciones aconsejaban utilizar modelos psicométricos que «permitan obtener la separabilidad de los parámetros de las personas y los calificadores» (Tesio, Simone, Grzeda, Ponzio, Dati, Zaratin, & Battaglia, 2015).

Los modelos de la teoría de la respuesta al ítem (TRI) permiten superar las limitaciones señaladas y generan una nueva tecnología psicométrica que complementa al modelo clásico de la teoría clásica de los tests (TCT). En ellos, es posible situar en un punto del espacio del rasgo o atributo tanto ítems como personas. La probabilidad de que un sujeto responda correctamente a un ítem está relacionado con la diferencia entre la capacidad del candidato y la dificultad del ítem. La colocación de las personas en el espacio del atributo depende de la cantidad que tengan de este, mientras que los ítems se sitúan dependiendo de la cantidad de rasgo que exijan para su correcta ejecución. Según este modelo, por tanto, el nivel de aptitud de un candidato es independiente del test aplicado y resulta posible representar el atributo objeto de la medición en una única dimensión en la que se sitúan conjuntamente personas e ítems.

Entre los modelos TRI, uno de los modelos dicotómicos más conocidos es el modelo de Rasch, que fue presentado en 1960 por el matemático danés Georg Rasch (1960/1980). Al conocer el nivel del candidato y la dificultad del ítem, es posible determinar la probabilidad de que una respuesta sea correcta (Prieto & Delgado, 2003), de tal forma que el nivel de aptitud de un candidato es independiente del test aplicado (Martínez, Hernández & Hernández, 2006, p. 130). Pero como este modelo únicamente resulta de utilidad para trabajar con ítems dicotómicos, se han desarrollado diversos modelos TRI que permiten analizar los datos que se obtienen en procesos de calificación mediante ítems politómicos como, por ejemplo, el modelo MFRM (Many Facet Rasch Measurement) (Linacre, 2012).

El modelo MFRM resulta adecuado para analizar de forma simultánea diferentes facetas que pueden tener un impacto relevante en los resultados de evaluación (Eckes, 2011, p. 12). Como en el resto de los modelos de Rasch, su principal característica es la invarianza de la medida, también llamada objetividad específica, y la suficiencia, que implica que la puntuación bruta obtenida por un examinado es el estadístico suficiente para estimar su parámetro en la escala *logit*. Asimismo, la suma de las calificaciones otorgadas por un calificador a un grupo de candidatos es el estadístico suficiente para estimar el parámetro del calificador en la escala de severidad. De este modo, y por medio del modelo, es posible obtener de manera independiente estimaciones en una misma escala de los diferentes parámetros de cada una de las facetas implicadas en la evaluación, que son las que pueden contribuir a la variabilidad de las medidas.

Para realizar los análisis con el modelo MFRM se utilizó el programa FACETS, en concreto la versión 3.70.0 (Linacre, 2012). FACETS permite estimar “los parámetros mediante el método de estimación conjunta por máxima verosimilitud (JML)” (Prieto, G., 2011, p. 234).

## 6 Análisis de las facetas

Cuando en un proceso de evaluación los calificadores evalúan un único atributo del desempeño del candidato en una tarea, es posible analizar al menos dos facetas: calificadores y candidatos. En caso de que los candidatos se enfrenten a varias tareas y los calificadores, en consecuencia, califiquen su actuación de manera independiente, es necesario tener en cuenta una tercera faceta: la tarea. Si las tres facetas están claramente definidas, la expresión formal del modelo MFRM es la siguiente (Eckes, 2011, p. 14):

$$\mathrm{In}\left[\frac{P_{nljk}}{P_{nljk-1}}\right] = \mathrm{B}_n - \mathrm{D}_l - \mathrm{R}_j - \mathrm{F}_k,$$

Donde,

$P_{nljk}$ es la probabilidad de que un candidato n reciba la calificación k en la tarea l por el calificador j.

$P_{nljk-1}$ es la probabilidad de que un candidato n reciba la calificación inferior (k-1) en la tarea l por el calificador j.

$\mathrm{B}_n$ es la competencia —valor de la variable latente— del candidato n.

$\mathrm{D}_l$ es la dificultad de la tarea l.

$\mathrm{R}_j$ es la severidad del calificador j.

$\mathrm{F}_k$ es la dificultad de recibir la calificación de k en relación con la categoría adyacente inferior k-1.

## 7 Análisis de la severidad de los calificadores

### 7.1 Sistema de doble calificación

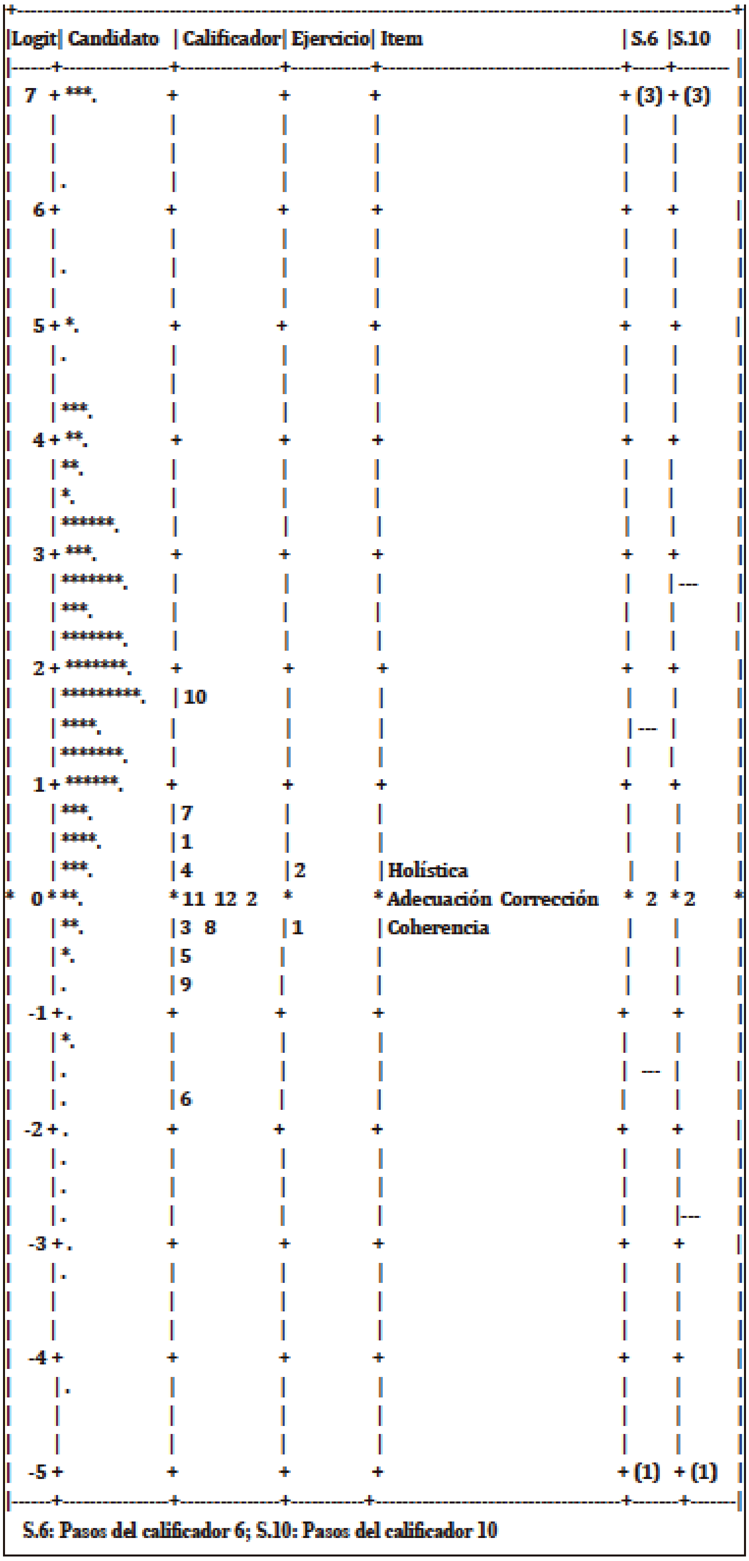
En la tabla 1 se muestran los datos que se obtuvieron en el proceso de calificación del 88,7% de los candidatos (n=3858) que se calificaron por medio del sistema de doble calificación. Las facetas estudiadas fueron las siguientes: 1 = candidato; 2 = calificador; 3 = tarea, y 4 = atributo.

El objetivo de este primer análisis es determinar la localización de las variables dentro del mapa. En la primera columna de la tabla 1, comenzando desde la izquierda, figura la escala *logit* en la que, como es habitual en los modelos Rasch, se suele situar el punto 0 en la dificultad media de las tareas, de los atributos y de la severidad media de los calificadores. Únicamente se permite variar libremente la faceta correspondiente a los examinados. Aunque teóricamente la escala *logit* puede adoptar valores entre 0 ± ∞, en la gran mayoría de los casos se sitúa en el rango ± 5 (Prieto & Delgado 2003, p. 95).

En la segunda de las columnas, la de candidatos, se muestra la distribución de esta faceta en la escala de *logit*. En las tablas, los asteriscos (*) representan frecuencias de candidatos. Cada asterisco representa a dos sujetos y cada punto (.) representa una frecuencia inferior. Los candidatos con mayor puntuación se sitúan en la parte superior de la tabla, mientras que los de menor puntuación se encuentran en la parte inferior.

En la columna *Calificador* se visualiza el mapa de los examinadores, y en la cuarta columna, la de *Tarea*, se muestra el nivel de dificultad de cada una de las tareas que integran las dos pruebas, ordenadas en la escala de *logit* de más difícil (arriba) a más fácil (abajo).

Las tareas, que han sido resueltas por todos los candidatos, está calibradas en la misma escala de intervalos (*logit*) que los candidatos. De este modo es posible comparar e interpretar los resultados de la competencia de los candidatos y la dificultad de las tareas en un mismo marco de referencia. Pero en el caso de la faceta calificador, debido al sistema de reparto utilizado, únicamente podemos comparar la situación dentro del mapa de la variable de los dos examinadores que han calificado los mismos exámenes (original y fotocopia), pero no es posible establecer ningún tipo de comparación con el resto. La ausencia de este elemento de conexión dificulta la posible comparación entre el equipo de examinadores.



```
+----------------------------------------------------------------------------------------------+
|Logit| Candidato     | Calificador| Ejercicio| Item                            | S.6 | S.10  |
|-----+---------------+-------------+----------+---------------------------------+-----+-------|
|  7  + ***.          +             +          +                                 + (3) + (3)   |
|     |               |             |          |                                 |     |       |
|     |               |             |          |                                 |     |       |
|     | .             |             |          |                                 |     |       |
|  6  +               +             +          +                                 +     +       |
|     |               |             |          |                                 |     |       |
|     | .             |             |          |                                 |     |       |
|     |               |             |          |                                 |     |       |
|  5  + *.            +             +          +                                 +     +       |
|     | .             |             |          |                                 |     |       |
|     |               |             |          |                                 |     |       |
|     | ***.          |             |          |                                 |     |       |
|  4  + **.           +             +          +                                 +     +       |
|     | **.           |             |          |                                 |     |       |
|     | *.            |             |          |                                 |     |       |
|     | ******.       |             |          |                                 |     |       |
|  3  + ***.          +             +          +                                 +     +       |
|     | *******.      |             |          |                                 |     | ---   |
|     | ***.          |             |          |                                 |     |       |
|     | *******.      |             |          |                                 |     |       |
|  2  + *******.      +             +          +                                 +     +       |
|     | *********.    | 10          |          |                                 |     |       |
|     | ****.         |             |          |                                 | --- |       |
|     | *******.      |             |          |                                 |     |       |
|  1  + ******.       +             +          +                                 +     +       |
|     | ***.          | 7           |          |                                 |     |       |
|     | ****.         | 1           |          |                                 |     |       |
|     | ***.          | 4           | 2        | Holística                       |     |       |
*  0  * **.           * 11 12 2     *          * Adecuación Corrección           *  2  * 2     *
|     | **.           | 3  8        | 1        | Coherencia                      |     |       |
|     | *.            | 5           |          |                                 |     |       |
|     | .             | 9           |          |                                 |     |       |
| -1  + .             +             +          +                                 +     +       |
|     | *.            |             |          |                                 |     |       |
|     | .             |             |          |                                 | --- |       |
|     | .             | 6           |          |                                 |     |       |
| -2  + .             +             +          +                                 +     +       |
|     | .             |             |          |                                 |     |       |
|     | .             |             |          |                                 |     |       |
|     | .             |             |          |                                 |     | ---   |
| -3  + .             +             +          +                                 +     +       |
|     | .             |             |          |                                 |     |       |
|     |               |             |          |                                 |     |       |
|     |               |             |          |                                 |     |       |
| -4  +               +             +          +                                 +     +       |
|     | .             |             |          |                                 |     |       |
|     |               |             |          |                                 |     |       |
|     |               |             |          |                                 |     |       |
| -5  +               +             +          +                                 + (1) + (1)   |
|-----+---------------+-------------+----------+---------------------------------+-----+-------|
  S.6: Pasos del calificador 6; S.10: Pasos del calificador 10
```

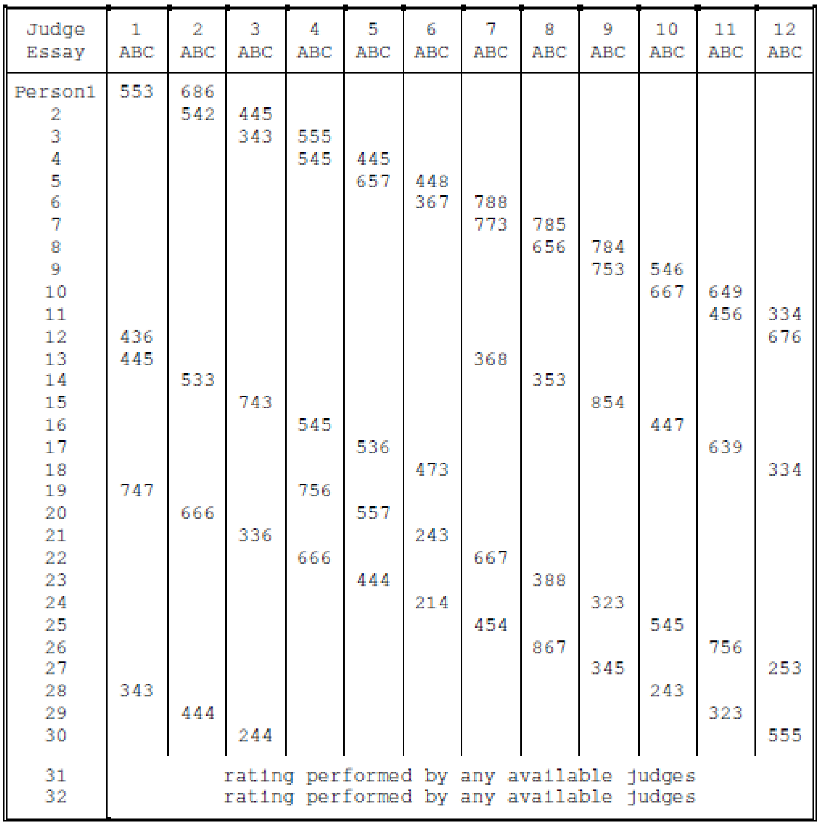
**Tabla 1.** Prieto, G (2011). Mapa de la variable con el procedimiento de doble calificación

### 7.2 Establecimiento de una red de calificadores

Para la calificación del restante 10,3% de candidatos (n=443,) se estableció una red con los doce calificadores que participaron en el proceso de calificación, de forma que todos compartieran pruebas con el resto de examinadores para que quedaran conectados calificadores, pruebas y candidatos. El proceso de reparto de exámenes y de asignación de los mismos se realizó mediante un sistema informático de visualización automatizada de pruebas.

La manera de organizar la distribución de las pruebas de los candidatos entre los calificadores resulta de especial relevancia. Teóricamente, si no existieran limitaciones de tiempo ni de horas de trabajo, lo ideal sería que todos los examinadores calificaran a todos los candidatos. De este modo, al hallar la media aritmética de las puntuaciones de los calificadores, se sabría cuál es su grado de severidad. Lamentablemente, cuando el número de candidatos es elevado esta propuesta resulta inviable.

En el mismo artículo, Linacre & Wright (2002, figura 2. Cf. también Eckes, 2011 y Tesio et al., 2015) sugieren un procedimiento que permite disminuir el número de calificaciones totales, de forma que el proceso sea viable y se establezca una red entre los parámetros implicados en el proceso: calificadores, candidatos y pruebas, de manera que todos queden relacionados entre sí. Aunque se elimina un número considerable de calificaciones, se mantiene la conexión entre candidatos, calificadores y pruebas, ya que al menos dos examinadores califican cada prueba y cada candidato comparte calificador con otro candidato. El ahorro de calificaciones que supone la aplicación de este primer proceso propuesto es del 83%, ya que únicamente es necesario realizar el 17% de las calificaciones que se harían si todos los examinadores calificaran a todos los candidatos (tabla 2). Lógicamente, este ahorro se consigue a costa de disminuir la precisión de las observaciones que se obtienen.



| Judge<br>Essay | 1<br>ABC | 2<br>ABC | 3<br>ABC | 4<br>ABC | 5<br>ABC | 6<br>ABC | 7<br>ABC | 8<br>ABC | 9<br>ABC | 10<br>ABC | 11<br>ABC | 12<br>ABC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Person1 | 553 | 686 | | | | | | | | | | |
| 2 | | 542 | 445 | | | | | | | | | |
| 3 | | | 343 | 555 | | | | | | | | |
| 4 | | | | 545 | 445 | | | | | | | |
| 5 | | | | | 657 | 448 | | | | | | |
| 6 | | | | | | 367 | 788 | | | | | |
| 7 | | | | | | | 773 | 785 | | | | |
| 8 | | | | | | | | 656 | 784 | | | |
| 9 | | | | | | | | | 753 | 546 | | |
| 10 | | | | | | | | | | 667 | 649 | |
| 11 | | | | | | | | | | | 456 | 334 |
| 12 | 436 | | | | | | | | | | | 676 |
| 13 | 445 | | | | | | 368 | | | | | |
| 14 | | 533 | | | | | | 353 | | | | |
| 15 | | | 743 | | | | | | 854 | | | |
| 16 | | | | 545 | | | | | | 447 | | |
| 17 | | | | | 536 | | | | | | 639 | |
| 18 | | | | | | 473 | | | | | | 334 |
| 19 | 747 | | | 756 | | | | | | | | |
| 20 | | 666 | | | 557 | | | | | | | |
| 21 | | | 336 | | | 243 | | | | | | |
| 22 | | | | 666 | | | 667 | | | | | |
| 23 | | | | | 444 | | | 388 | | | | |
| 24 | | | | | | 214 | | | 323 | | | |
| 25 | | | | | | | 454 | | | 545 | | |
| 26 | | | | | | | | 867 | | | 756 | |
| 27 | | | | | | | | | 345 | | | 253 |
| 28 | 343 | | | | | | | | | 243 | | |
| 29 | | 444 | | | | | | | | | 323 | |
| 30 | | | 244 | | | | | | | | | 555 |
| 31 | rating performed by any available judges | | | | | | | | | | | |
| 32 | rating performed by any available judges | | | | | | | | | | | |

**Tabla 2**. Linacre y Wright (2002). Procedimiento de reducción del número de calificaciones

En nuestro trabajo seguimos el procedimiento descrito arriba aunque ligeramente modificado:

| Calificador | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Candidato 1 | x | x | | | | | | | | | | |
| 2 | | x | x | | | | | | | | | |
| 3 | | | x | x | | | | | | | | |
| 4 | | | | x | x | | | | | | | |
| 5 | | | | | x | x | | | | | | |
| 6 | | | | | | x | x | | | | | |
| 7 | | | | | | | x | x | | | | |
| 8 | | | | | | | | x | x | | | |
| 9 | | | | | | | | | x | x | | |
| 10 | | | | | | | | | | x | x | |
| 11 | | | | | | | | | | | x | x |
| 12 | x | | x | | | | | | | | | |
| 13 | | x | | x | | | | | | | | |
| 14 | | | x | | x | | | | | | | |
| 15 | | | | x | | x | | | | | | |
| 16 | | | | | x | | x | | | | | |
| 17 | | | | | | x | | x | | | | |
| 18 | | | | | | | x | | x | | | |
| 19 | | | | | | | | x | | x | | |
| 20 | | | | | | | | | x | | x | |
| 21 | | | | | | | | | | x | | x |
| 22 | x | | | x | | | | | | | | |
| 23 | | x | | | x | | | | | | | |
| 24 | | | x | | | x | | | | | | |
| 25 | | | | x | | | x | | | | | |
| 26 | | | | | x | | | x | | | | |
| 27 | | | | | | x | | | x | | | |
| 28 | | | | | | | x | | | x | | |
| 29 | | | | | | | | x | | | x | |
| 30 | | | | | | | | | x | | | x |
| 31 | x | | | | x | | | | | | | |
| 32 | | x | | | | x | | | | | | |
| 33 | | | x | | | | x | | | | | |
| 34 | | | | x | | | | x | | | | |
| 35 | | | | | x | | | | x | | | |
| 36 | | | | | | x | | | | x | | |
| 37 | | | | | | | x | | | | x | |
| 38 | | | | | | | | x | | | | x |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 39 | x | | | | | x | | | | | | |
| 40 | | x | | | | | x | | | | | |
| 41 | | | x | | | | | x | | | | |
| 42 | | | | x | | | | | x | | | |
| 43 | | | | | x | | | | | x | | |
| 44 | | | | | | x | | | | | x | |
| 45 | | | | | | | x | | | | | x |
| 46 | x | | | | | | x | | | | | |
| 47 | | x | | | | | | x | | | | |
| 48 | | | x | | | | | | x | | | |
| 49 | | | | x | | | | | | x | | |
| 50 | | | | | x | | | | | | x | |
| 51 | | | | | | x | | | | | | x |
| 52 | x | | | | | | | x | | | | |
| 53 | | x | | | | | | | x | | | |
| 54 | | | x | | | | | | | x | | |
| 55 | | | | x | | | | | | | x | |
| 56 | | | | | x | | | | | | | x |
| 57 | x | | | | | | | | x | | | |
| 58 | | x | | | | | | | | x | | |
| 59 | | | x | | | | | | | | x | |
| 60 | | | | x | | | | | | | | x |
| 61 | x | | | | | | | | | x | | |
| 62 | | x | | | | | | | | | x | |
| 63 | | | x | | | | | | | | | x |
| 64 | x | | | | | | | | | | x | |
| 65 | | x | | | | | | | | | | x |
| 66 | x | | | | | | | | | | | x |

**Tabla 3**. Procedimiento de reparto para el establecimiento de una red de calificadores

De este modo, es posible visualizar el mapa de la variable en una única tabla en la que se presentan los elementos de las diferentes facetas analizadas calibrados en la misma escala de intervalos (*logit*) y se pueden comparar e interpretar los resultados de la competencia de los candidatos, la dificultad tanto de las tareas como de los atributos, la localización de los valores de paso de los valores adyacentes en un mismo marco de referencia, así como la severidad de los calificadores (tabla 4).

## 8 Conclusión

El objetivo de este trabajo es analizar la severidad de un equipo de examinadores que han participado en el proceso de calificación de una prueba de desempeño. Para garantizar la conectividad entre todos ellos de forma que el proceso de calificación fuera viable, seguimos el procedimiento propuesto por Linacre y Wright en 2002, ligeramente modificado, de rotación de los examinados y los calificadores, según el cual cada candidato comparte calificador con otro candidato y al menos dos calificadores evalúan cada prueba. De este modo fue posible comparar el nivel de severidad del equipo de calificadores durante el proceso de calificación.

```
+-------------------------------------------------------------------------------+
|Measr|+Candidato|-Calificador|-Tarea   |-Atributo                        |Scale|
|-----+----------+------------+---------+---------------------------------+-----|
|   9 + *        +            +         +                                 + (3) |
|     |          |            |         |                                 |     |
|     | .        |            |         |                                 |     |
|   8 + .        +            +         +                                 +     |
|     | .        |            |         |                                 |     |
|     | .        |            |         |                                 |     |
|   7 + .        +            +         +                                 +     |
|     | .        |            |         |                                 |     |
|     | *.       |            |         |                                 |     |
|   6 + *.       +            +         +                                 +     |
|     | *.       |            |         |                                 |     |
|     | **.      |            |         |                                 |     |
|   5 + ***.     +            +         +                                 + --- |
|     | ***.     |            |         |                                 |     |
|     | **.      |            |         |                                 |     |
|   4 + *****    +            +         +                                 +     |
|     | *******. |            |         |                                 |     |
|     | ****.    |            |         |                                 |     |
|   3 + ******.  +            +         +                                 +     |
|     | ****.    |            |         |                                 |     |
|     | ****.    |            |         |                                 |  2  |
|   2 + ****.    +            +         +                                 +     |
|     | **       | 12         |         |                                 |     |
|     | ****.    | 10  7      |         |                                 |     |
|   1 + ****     +            +         +                                 +     |
|     | **.      | 2          |         |                                 |     |
|     | **.      | 8   9      | Tarea 3 |                                 |     |
*   0 * **.      * 11  5      * Tarea 2 * Adec-Coh   Corr-Alc   Holistica *     *
|     | *        | 4          | Tarea 1 |                                 | --- |
|     | **.      |            |         |                                 |     |
|  -1 + **.      + 1          +         +                                 +     |
|     | *.       |            |         |                                 |     |
|     | .        |            |         |                                 |     |
|  -2 + *        + 3          +         +                                 +     |
|     | .        | 6          |         |                                 |  1  |
|     | *        |            |         |                                 |     |
|  -3 + .        +            +         +                                 +     |
|     | .        |            |         |                                 |     |
|     |          |            |         |                                 |     |
|  -4 + .        +            +         +                                 +     |
|     |          |            |         |                                 | --- |
|     |          |            |         |                                 |     |
|  -5 + .        +            +         +                                 + (0) |
+-------------------------------------------------------------------------------+
 Measr = Medición Rasch (logit)
 Scale = Escala en puntuaciones directas (0-3)
```

**Tabla 4**. Mapa de la variable con los calificadores en red

En la tabla 4 sí fue posible comparar el nivel de severidad/benignidad de todos los calificadores entre sí. Los más severos fueron el 12, el 10 y el 7, mientras que los más benévolos fueron el 6, el 3 y el 1. En la franja central (no son ni excesivamente severos ni demasiado benévolos) se encuentran los calificadores 4, 11, 5, 8, 9  2.

## Referencias bibligráficas

Cronbach, L. J. (1990). *Essentials of Psychological Testing*. Nueva York: Harper & Row.

Eckes, T. (2011). *Introduction to Many-facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments*. Frankfurt: Peter Lang.

Guilford, J. P. (1954). *Psychometric Methods*. New York: McGraw-Hill.

Linacre, J. M. (2012). *Facets Computer Program for Many-facet Rasch Measurement, versión 3.70.0*. Beaverton: Winsteps.com.

Linacre, J. M. & Wright, B. D. (2002). Construction of measures from Many-facet data. *Journal of Applied Measurement, 3*(4), 484–509.

Martínez, Mª. R. (2010). La evaluación del desempeño, *Papeles del Psicólogo, 31*(1), 85–96.

Mártínez, Mª. R., Hernández, Mª. J., & Hernández, Mª. V. (2006). *Psicometría*. Madrid: Alianza Editorial.

Myford, C. M. and Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. In E. V. Smith,& R. M. Smith (Eds.). *Introduction to Rasch measurement* (pp. 460–517). Maple Grove: JAM Press.

Popham, W. J. (1990). *Problemas y técnicas de la evaluación educativa*. Madrid: Anaya.

Prieto, G. (2011). Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement. *Psicothema, 23*(2), 233–238.

Prieto, G. & Delgado, A. R. (2003). Análisis de un test mediante el modelo de Rasch. *Phicothema*, *15*, 1, 94–100.

Prieto, J. M. (2016). *Estudio del comportamiento de los examinadores de la prueba de expresión escrita mediante el modelo Many-Facet Rasch Measurement (MFRM) en el contexto de un examen de dominio: el diploma de español nivel A2* <http://hdl.handle.net/10366/128550> (7 de septiembre de 2017).

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research/Chicago: University of Chicago Press.

Tesio, L.; Simone, A; Grzeda M. T., Ponzio, M., Dati, G, Zaratin, P, & Battaglia, M. A. (2015). Funding Medical Research Projects: Taking into Account Referees' Severity and Consistency through Many-Faceted Rasch Modeling of Projects' Scores. *Journal of Applied Measurement*, *16*(2), 129–152.

# Predicting Readability of Texts for Italian L2 Students: A Preliminary Study

**Giuliana Grego Bolli**, CVCL (Centre for Language Assessment and Certification) – University for Foreigners of Perugia, Italy
**Danilo Rini**, CVCL (Centre for Language Assessment and Certification) – University for Foreigners of Perugia, Italy
**Stefania Spina**, Department of Human and Social Sciences – University for Foreigners of Perugia, Italy

**Abstract:** Text selection and comparability for L2 students to read and comprehend are central concerns both for teaching and assessment purposes.Compared to subjective selection. quantitative approaches provide more objective information, analysing texts at language and discourse level (Khalifa & Weir, 2009). Readability formulae such as the Flesch Reading Ease, the Flesch-Kincaid Grade Level and, for Italian, the GulpEase index (Lucisano and Piemontese, 1988), do not fully addressed the issue of text complexity. A new readability formula called Coh-Metrix was proposed (Crossley, Geenfield, & McNamara 2008), which takes into account a wider set of language and discourse features. A similar approach was proposed to assess readability of Italian texts through a tool called READ-IT (Dell'Orletta, Montemagni, & Venturi 2011). While READ-IT was tested on newspaper texts randomly selected, this contribution focuses on the development of a similar computational tool applied on texts specifically selected in the context of assessing Italian as L2. Two text corpora have been collected from the CELI (Certificates of Italian Language) item bank at B2 and C2 level. Statistical differences in the occurrence of a set of linguistic and discursive features have been analysed according to four different categories: length features, lexical features, morpho-syntactic features, and discursive features.

## 1 Introduction

The selection and the level of difficulty of texts to read is one of the central issues both for teachers and language testers. In the context of assessment, the decision taken with regard to texts in the case of reading tests has serious implication in the interpretation of test scores, hence in providing validity evidence to the overall testing process.

Focusing on reading comprehension, texts are mostly subjectively selected by experienced teachers and test producers depending on several aspects: specific curricula, programmes, guidelines and test specifications.

Other aspects such as the definition of readers' population, their linguistic needs, their educational background, their age, consequently involving other aspects such as text genre, text type, tasks to be assigned, are also taken into account.

There is quite a wide consensus in the literature about a set of other characteristics that can have an impact on the difficulty of a reading comprehension test, also in terms of cognitive demands imposed upon the reader (Bachman & Palmer, 2010; Purpura, 2014). These characteristics can be also measurable or judged by competent teachers or test developers, as often happens. They are: text length, grammatical complexity, word frequency, cohesion, rhetorical organisation, genre, text abstractness, subject knowledge and cultural knowledge. All these aspects relates to readability, which means to find measures of text's ease or difficulty in terms of comprehension (Green, Ünaldi & Weir, 2010; Khalifa & Weir, 2009).

Both qualitative and quantitative analysis can strongly contribute to a more comprehensive, evidence based approach to readability and hence on selecting and scaling texts in terms of difficulty both for assessment and teaching purposes. This kind of support is not provided by the Common European Framework of Reference for Languages (CEFR) descriptors and scales related to reading comprehension: they can provide information supporting text selection, but not in terms of readability.

Also within the CELI (Certificates of Italian Language) certification system, produced by CVCL (Centre for Language Assessment and Certification) at the University for Foreigners of Perugia, texts selection has been so far based on this set of characteristics subjectively assessed by CVCL experts' informed analysis.

Bearing in mind that in language testing terms, the decision taken with regard to texts in the case of reading tests may affect the interpretability of score outcomes, it is unquestionable that quantitative approaches, supported by automated analysis and systematic data collection, can provide more objective information, analysing texts on multiple levels of language and discourse and providing test producers and item writers ways to evaluate this aspect of test validity.

It is well known that readability assessment has been a central research topic for the past 80 years. The development of quantitative tools, such as Flesch Reading Ease, the Flesch-Kincaid Grade Level and, for Italian, the GulpEase index (Lucisano & Piemontese, 1988), opened the way to an automated textual description providing a more evidence-based approach to text selection and scaling.

Over the last decades, the automatic assessment of readability has received increasing attention: advances in computational linguistics and development of corpora, jointly with the availability of sophisticated language technologies, allow the capuring of a wide variety of more and more complex linguistic features affecting the readability of a text.

More recently, particularly in the last 20 years, scientific investigation of reading also benefited from more complex and automated measures of text characteristics, and systematic data collection, such as Coh-Metrix (Graesser, McNamara & Kulikowich, 2011; Graesser, McNamara, Louwerse & Cai, 2004) were proposed.

Taking all this into account, this paper reports on the development, at the University for Foreigners of Perugia, of a similar computational tool applied to texts specifically selected in the context of the CELI examinations suite. The tool itself and the consequent data collection and analysis will give more information and evidence about text readability as a part of the continuous validation process applied in the context of CELI.

## 2 Selection of texts and tasks in CELI 3 (B2) and in CELI 5 (C2)

In the routine work of texts selection for the Reading component of CELI exams, the following basic aspects are taken into account by CVCL item writers: the characteristics of texts as shown in the CEFR, and the genres identified in the Profilo della Lingua Italiana (Spinelli & Parizzi, 2008). A detailed overview of CELI exams specifications can be found in Grego Bolli & Spiti (2004).

Amongst the CEFR descriptors concerning Reading skills used in the selection of texts, “Overall reading comprehension”, “Reading for orientation” and “Reading for information and argument” can be found. Along with them, the Profilo helps in identifying the genres and text types. It has to be underlined how the Profilo does not include any referential for C2 level, but, on the other hand, C2 level language users can deal with any type of textual genre. With these indications in mind, in CELI exams the text types used for assessment of language competence for CELI 3 (B2) include fiction and non-fiction books, magazine and newspaper articles, textbooks, interviews, and personal letters, whereas for CELI 5 (C2), fiction and non-fiction books, including literary journals, specialist magazines, newspapers, textbooks and essays, personal letters, regulations, memoranda, reports and papers are used.

Tasks in the exam papers have the objective of testing the following sub-skills, for CELI 3: reading for gist; identifying point of view; identifying main points; reading for detailed information, skimming and scanning; and for CELI 5, along with the above mentioned: identifying point of view and tone, guessing meaning from context, recognising the organization of a text, reading for detailed information. The length of texts used for testing reading skills vary from 250–350 words in CELI 3  to 600–650 in CELI 5, and in both papers the answer format for the texts taken into account include 4-option multiple-choice, and short answers. It has to be added that items are generally calibrated according to IRT model based on Rasch analysis, placing the item difficulty at the pre-established level.

## 3 Method

The main question we are trying to answer in this study is: how can we operationalise complexity in order to measure it in texts to be selected for learning and assessment purposes?

From the theoretical point of view, we considered two different models in the field of measures of complexity.

The first model is Coh-Metrix (Graesser et al., 2004): it is a well-established project that takes into account a wide set of language and discourse features, based on 108 indices. While these indices belong to different levels of linguistic analysis, Coh-Metrix is mainly focused on cohesion, and is specifically targeted to English texts.

The second model is READ-IT (Dell’Orletta et al., 2011), which is targeted to Italian texts, and aimed at text simplification: its intended audience are mainly people with low literacy skills and/or with cognitive impairment. In contrast with Coh-Metrix, READ-IT is mainly focused on lexical and syntactic features, such as syntactic dependencies or part-of-speech probability.

None of these two models is sufficient to achieve the goal of developing a computational tool to be used with texts specifically selected in the context of learning and assessing Italian as an L2. The methodology we followed was based on two different steps: the corpus-based feature selection process, and the tool creation and testing.

The first task we had to perform was the identification of a set of linguistic features to be used in order to establish text difficulty. As we still are at an early stage of the project, in the features selection process we preferred easy-to-identify features which could be reliably detected within the output of computational resources.

To this aim, we collected two corpora of texts from the CELI item bank at B2 and C2 level. It is important to stress that these texts were selected and assigned to a specific level by experienced, professional teachers. With the 213 selected texts, we built a corpus with 133,364 tokens (B2 level: 122 texts and 59,423 tokens; C2 level: 91 texts and 73,941 tokens), which was xml-annotated and post-tagged (the tag-set and annotation scheme were the same as those used for the annotation of a reference Italian corpus; see Spina, 2014).

As complexity is intrinsically multifactorial, we selected a wide set of linguistic and discursive features, that, in our opinion, affect texts comprehension and systematically vary as a function of types of texts and grade level. In addition, these features show a growing computational complexity, so as to follow the different levels of linguistic analysis automatically carried out on texts.

The selected linguistic features are distributed in the following four categories:

- raw-text features (length features)
- lexical features
- morpho-syntactic features
- discursive features.

## 4 Results

### 4.1 Raw-text features

Raw-text features are from the computational point of view the simplest category, and were typically used within traditional readability metrics. Nevertheless, they can give a contribution in predicting text complexity: higher level texts (C2) are formed by longer sentences (B2: 18.1; C2: 20.8 words per sentence), and by slightly longer words (B2: 4.8; C2: 5 mean word length).

### 4.2 Lexical features

We selected four different lexical matrix that are generally considered in the computation of linguistic complexity: lexical diversity (Aluisio, Specia, Gasperin, & Scarton, 2010), lexical density (Feng, Elhadad & Huenerfauth, 2009), basic Italian vocabulary rate (Dell’Orletta et al., 2011), and the percentage of concrete/abstract nouns.

Lexical diversity (Malvern, Richards, Chipere & Durán, 2004), defined as the ratio of total number of words to the number of different unique words, is a measure of the amount of different words used in a text. A text with a higher score of lexical diversity includes more different words, and is therefore more complex, while texts with lower scores tend to repeat the same words many times. We used the Guiraud index (Guiraud, 1954) as an index of lexical diversity. This index was used instead of type/token ratio because it compensates the systematic decrease of the number of tokens when texts to compare have different lengths (e.g. Van Hout & Vermeer, 2007). The respective values of lexical diversity (B2: 43.3; C2: 51.9) show that higher level texts tend to include more different words, and, as a consequence, to be more complex.

Lexical density (Ure, 1971) refers to the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of words in a text. The idea behind this measure is that more dense texts are also more difficult to understand. The respective values (B2: 44.7; C2: 45.8) show that lexical density also contribute to the greater complexity of C2 texts.

The basic Italian vocabulary rate measures the internal composition of the vocabulary of the texts. To this end, we took as a reference the New Basic Italian Vocabulary (NVdB) by De Mauro & Chiari (forthcoming), that includes a list of 7,500 words highly familiar to native speakers of Italian. In more detail, we calculated two different features corresponding to: a) the percentage of lemmas on this reference list that are used in the texts; b) the internal distribution of the occurring basic Italian vocabulary words, and in particular the 2,000 most frequent, or fundamental words. Both the 7,500 total words (B2: 77.9; C2: 76.4) and the 2,000 fundamental words (B2: 71.2; C2: 68.8) are used more in the easier texts. This reveals, hence, a greater use of more frequent, and thus easier, lexical items in lower level texts.

Finally, we considered the use of concrete and abstract nouns. The percentage of concrete nouns is significantly higher in B2 texts (B2: 56.7; C2: 48), while abstract nouns are used more in C2 texts (B2: 11.8; C2: 18.3). This finding is relevant for our research, because concrete nouns are more familiar and then easier for the reader, as familiarity has a strong impact on a wide range of cognitive processes, including comprehension.

### 4.3 Morpho-syntactic features

In general, the morpho-syntactic features selected for this study seemed to affect texts complexity less than other linguistic features: we did not find significant differences in part-of-speech distribution and in the global number of subordinate clauses, although subordination is traditionally acknowledged as an index of structural linguistic complexity. The only kind of subordinate clause that is used significantly more in C2 texts is relative (log-likelihood = 13.40).

### 4.4 Discursive features

We believe that cohesion plays a key role in text readability. By cohesion we refer to the “characteristics of the explicit text that play some role in helping the reader mentally connect ideas in the text" (Graesser, McNamara, & Louwerse, 2003).

Following the Coh-metrix model, we studied two different dimensions of cohesion: the referential cohesion and the deep cohesion.

Referential cohesion can be measured by assessing the overlap between adjacent sentences: high cohesion texts contain words that overlap across sentences, forming threads that help readers to recover the message, while low cohesion texts have to count on knowledge-based inferences to fill the gaps.

What we found in our data was that adjacent sentences that contain overlapping nouns are significantly more frequent in B2, easier texts.

The following example shows the use of the overlapping noun medico (“physician”) accross three adjacent sentences.

> Conosco medici laureati con 110 e lode da cui non mi farei curare nemmeno un'unghia. Ho fiducia in questo medico falso. Non lo cambierei con nessun altro medico.

Moving to deep cohesion, taking for granted that cohesion gaps increase reading time and complexity, we measured the use of connectives, which play an important role in the creation of logical relations within text meanings, and provide clues about text organisation (Halliday & Hasan, 1976).

Based on eight classes of connectives (causal, temporal, additive, adversative, marking results, transitions, alternative or reformulation/specification), we found that in some cases, as in causal connectives, there is a substantial equivalence in the two levels of texts, but in other cases, as in temporal connectives, there is a significant difference, and connectives are much more frequent in easier texts.

## 5 Conclusions

We presented an exploratory study on the possibility of measuring complexity in Italian texts, selected for L2 learning and assessment purposes.

The process of corpus-based feature selection, resulting in four dimensions with growing computational complexity, revealed significant differences in texts assigned to specific CEFR levels by experienced teachers. These differences emerged particularly in lexical and discursive features. This analysis also confirmed that the use of a quantitative approach should always be part of the cyclic process of text selection.

Future work will be needed in order to fulfil the aim of creating a tool for the automatic assessment of complexity. One future direction will be the refinement of the linguistic indices of complexity, with a deeper analysis of overlap across sentences, and the addition of narrativity, which is a major predictor of text complexity.

## References

Aluisio, S., Specia, L., Gasperin, C., & Scarton, C. (2010). *Readability assessment for text simplification*. Paper presented at the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications.

Bachman, L. F. & Palmer, A. , S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Crossley, S.A., Greenfield, J., & McNamara, D.S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475–493.

De Mauro, T., & Chiari, I. (forthcoming). *Il Nuovo Vocabolario di Base della Lingua Italiana*.

Dell'Orletta, F., Montemagni, S., & Venturi, G. (2011). *READ–IT: Assessing readability of Italian texts with a view to text simplification*. Paper presented at the 2nd Workshop on Speech and Language Processing for Assistive Technologies, Edinburgh.

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). *Cognitively motivated features for readability assessment*. Paper presented at the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09).

Green, A., Ünaldi, A., & Weir, C. J. (2010). Empiricism versus connoisseurship: establishing the appropriacy of text for testing reading for academic purposes. *Language Testing, 27*(3), 1–21.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet & C.E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82–98). New York: Guilford Publications.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, And Computers, 36*, 193–202.

Grego Bolli, G. & Spiti, M. G. (2004). *Misurare e valutare nella certificazione CELI*. Perugia: Edizioni Guerra.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Khalifa, H. & Weir, C. J. (2009). *Examining reading: research and practice in assessing second language reading*. Cambridge: UCLES/Cambridge University Press.

Lucisano, P. & Piemontese, M.E. (1988). GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città, 31*(3), 110–124.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development. Quantification and assessment*. London: Palgrave Macmillan.

Purpura, J. E. (2014). Cognition and language assessment. In Kunnan, A., J. (Ed.). *The Companion to Language Assessment volume III* (pp. 1,453–1,476). Oxford: Wiley Blackwell.

Spina S. (2014). Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In R. Basili, A. Lenci, & B. Magnini (Eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014* (354–359). Pisa: Pisa University Press.

Spinelli, B. & Parizzi, F. (2010). *Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2*. Firenze: La Nuova Italia.

Ure, J. (1971), Lexical density and register differentiation. In G. Perren & J. L. M. Trim (Eds.), *Applications of Linguistics. Selected Papers of the Second World Congress of Applied Linguistics* (pp. 443–452). Cambridge: Cambridge University Press.

Van Hout, R. & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 93–115). Cambridge: Cambridge University Press.

# The Impact of On-line Teaching Practices on Greek EFL Learners' Reading Perceptions & Exam Performance

**Trisevgeni Liontou**, Greek Ministry of Education, Greece

**Abstract:** This paper reports on a 1-year longitudinal study that adopted a blended teaching approach based on designing and implementing an online EFL course to be used by Greek students aged 12-15 years old alongside their more traditional face-to-face lessons. The reason for creating a more dynamic learning environment aligned with the rest of the curriculum was to increase EFL learners' engagement and motivation through their exposure to authentic online material and participation in a variety of reading, writing, speaking and listening tasks. To this end, a number of online activities were designed including: a) an online classroom with handouts, extra activities, resources and discussion groups for students to further develop their digital literacy along with their English language competence, b) a wiki for students to make a contribution and post their own messages on a specific topic, c) a series of Skype group discussions with invited external guest speakers, d) a private YouTube space for students to upload their videos and watch relevant EFL material. Data analysis of pre- and post-achievement tests on English language reading comprehension performance along with students' Computer-Assisted Language Learning (CALL) Attitude questionnaire showed that, in general, participants in this study had a positive attitude toward CALL while, at the same time, open online access technologies gave them the opportunity to further develop their EFL reading comprehension skills. The paper concludes by highlighting the fact that online class components were not designed around the tools, attempting to fit the online tools into a task-based EFL lesson, but rather served the learning objectives of the actual lesson based on a blended teaching approach, in which face-to-face and online learning activities were relevant to and complemented one another.

## 1 Introduction

According to McIntyre, Mirriahi, & Watson (2014, p. 2), "the Internet has significantly changed how we communicate with one another as well as how we access, share and facilitate information". Providing materials for students to complete courses online has created a new era for teaching, since not only can students benefit from collaborative learning but institutions and instructors can efficiently distribute materials and information (Levinsen, 2006; McIntyre et al., 2014; Parker, Maor, & Herrington, 2013). Such an approach is in line with integrative CALL based on "a perspective which seeks both to integrate various skills (e.g., listening, speaking, reading, and writing) and also integrate technology more fully into the language learning process" (Warschauer & Healey, 1998, p. 58). Due to its reported positive effect on language learning, the use of technology as a language acquisition medium has increased phenomenally in the last two decades (Greenfield, 2003). As Furstenberg (1997) notes, CALL is a tool that enhances learner-learner interaction, while Warschauer & Healey (1998) point out that CALL can help learners use language in authentic situations. In a similar line, Kelm (1998) also argues that CALL can help learners use language in authentic situations while promoting socialisation and communication among them.

Nevertheless, there seem to be a certain degree of resistance against the integration of CALL into EFL curricula "since some people may have negative attitude toward CALL because they think that it is a kind of unwanted 'luxurious' change" (Bulut & AbuSeileek, 2006, p. 15). To address concerns on the integration of CALL into ESL/EFL curricula, Gillespie and McKee (1999) suggest it is necessary to judge the success of CALL by investigating, amongst other things, students' attitude toward its effectiveness. Lasagabaster and Sierra (2003) also express the belief that researchers should take students' opinions into consideration when CALL programs are evaluated, since students are potential contributors to the development of their language learning tools. Based on the above literature and on the importance of focusing on understanding effective pedagogical strategies for online teaching from EFL learners' point of view, the aim of the present research is to empirically investigate the impact of online teaching practices on young intermediate EFL learners' motivation and reading comprehension competence through their participation in an online English language classroom.

This study bears resemblance to prior studies concerned with how CALL affects student achievement while investigating EFL learners' general attitudes toward computers and, more specifically, toward the use of computers when developing their reading comprehension skills. At the same time, the originality of the study lies in the combined purpose of identifying the relationship between EFL learners' attitude toward CALL and their level of achievement in EFL reading comprehension competence when adopting a blended task-based English language learning approach with young learners.

## 2 The study

### 2.1 Research aims

The main aim was to create a more dynamic learning environment aligned with the rest of the curriculum in order to increase young EFL learners' engagement and motivation through their exposure to authentic material and participation in real-life tasks. In accordance with the aims of the study, the following research questions were formulated:

(1) What is the general attitude of intermediate Greek learners of English towards the use of CALL in their language lessons?

(2) What is their attitude towards using CALL to enhance their EFL reading comprehension competence?

(3) Is there a significant improvement in intermediate EFL learners' reading comprehension competence after attending a 1-year online EFL reading course?

### 2.2 Objectives and design

The current study, which lasted one year and consisted of two face-to-face lessons per week plus online activities, was based on designing and implementing an online EFL course directed to a selected sample of 40 intermediate EFL students aged 12–15 years old alongside their more traditional face-to-face lessons.

### 2.3 Participants

Selected participants (N = 40) came from a junior high school located in Athens, Greece and had all been taught Information Technology as a compulsory school subject for five years before taking part in the study. As a part of their IT courses, students had been exposed to various word-processing and desktop publishing software applications and were familiar with online environments including wikis and YouTube. Participants were chosen for their high grades achieved in their IT school exams and were, therefore, expected to have a similar level of digital literacy. Their language proficiency (intermediate level-B1) was diagnosed through a calibrated English language test (*Cambridge English: Preliminary* – PET).

### 2.4 Tools and procedure

The online class components consisted of a free online Omnium classroom with online handouts, extra activities, text resources and discussion groups for students to further develop their digital literacy along with their English language competence. The OmniumClass is a free e-learning software package, designed to help teachers to quickly set up their online classes. Following parents' written consent, the intermediate EFL students taking part in the present study were able to perform different activities as registered users with controlled access. These included revising information presented in the classroom (handouts and video lectures), doing extra online activities such as computer-based quizzes with gap-filling, multiple-choice, true-false, drag-and-drop activities, accessing online resources such as e-books and electronic dictionaries, adding comments/suggestions/ideas for projects, topics they would like to talk about in-class, as well as posting their wikis on a variety of eating topics. Their individual contributions to each specific wiki formed part of their classroom evaluation so students were more than willing to post comments and share thoughts and knowledge with their classmates.

In addition, Skype was used to set up a series of guided group meetings with invited external guest speakers, including a dietician, a chef, a doctor and a gymnast, who contributed by discussing different eating-related topics with students.

Flickr was also used to allow students to upload their own projects and photos to the English Classroom gallery. Finally, students were asked to create their own "healthy eating" videos and post them in a private YouTube space shared only with their classmates. Through the use of video analytics in YouTube, it was possible to identify patterns of how students accessed and watched relevant material and further worked on their language skills.

On the other hand, data collection tools were used to gather valuable information on intermediate EFL learners' perceptions of online teaching practices. A 5-point Likert scale paper-and-pencil attitude questionnaire was administered to them upon completion of the course. To facilitate respondents' understanding and ease their answers, the questionnaire was written in respondents' native language, Greek. This minimised reliability and validity problems caused by the language factor. Participants were requested to rate their agreement or disagreement with 20 statements using a 5-point scale. Statements were related to their attitudes to online teaching practices, feelings of preference, enjoyment and motivation when taking part in online activities, as well as perceived difficulties encountered during the course. The CALL attitude questionnaire (see Appendix 1) used in the present study was an adapted version of the one used by Bulut and AbuSeileek (2006).

Furthermore, a standardised multiple-choice reading test was used to investigate the impact, if any, of online teaching practices on EFL students' reading skills. The test consisted of 4 texts with 5 multiple-choice reading comprehension questions per text and was administered to all participants at the beginning of the course. A parallel version of the same test was used to assess reading competence upon completion of the course. A total of 40 reading comprehension questions per student and 1,600 for the whole group of participants was collected. Once the questionnaires were collected, data was tabulated and synthesized for statistical analyses. Data coding consisted of assigning a code number to each item. Frequency distributions were then calculated. All percentages were reported as valid percentages with missing data excluded. The mean, median and standard deviation estimates were then used to indicate average responses and variability of attitudes. As Wiersma (2008) explained, survey results typically include this kind of descriptive information, since such an approach enables the researcher to provide general information about respondents' central tendency when answering each question, and further show how responses disperse around the centre. Finally, data were subjected to further statistical analysis using IBM SPSS 20.0 statistical package. As far as reading comprehension performance is concerned, the mean task scores per text of the 40 EFL learners were estimated. These mean scores, related to the specific multiple-choice reading comprehension questions included in each set of analysed texts, revealed significant relationships between mean reading performance before and after taking part in the designed online course.

## 3 Results and discussion

### 3.1 English language reading comprehension achievement tests

Data analysis of pre- and post-achievement tests of English reading comprehension was based on a total of 800 multiple-choice reading comprehension questions from the pre-test and 800 multiple-choice reading comprehension questions from the post-test. Results indicated that open online access technologies gave intermediate EFL participants the opportunity to enhance their reading skills through their exposure to authentic online material that did not form part of their traditional classroom-based English language lessons. More specifically, in order to compare the mean reading performance in the pre-test and post-test, a set of independent sample t-tests were carried out. The results of this analysis showed that EFL participants’ mean reading performance was significantly higher in the post-test which was parallel in form and level of difficulty to the pre-test, which could be partly attributed to their exposure to a wider range of online text resources and reading activities (t = 8.851, df = 38, p = .021).

### 3.2 Student attitude towards online EFL reading classes

In order to identify learners’ attitudes towards the use of online classes for the development of their reading comprehension competence, 5 related statements (Statements 16–20) were included in the questionnaire. As demonstrated in Table 1, the highest frequency score was 70% (Agree) for Statement 16: It is easy to access the meaning of words (e.g., use online dictionaries, pictures) to help me understand what I read in my online EFL classes, and for Statement 19: Reading via computers is more interesting when supported with visual information (Strongly Agree: 50%). These findings can be partly attributed to the fact that, since online reading classes included annotated texts and electronic dictionary use, students had the opportunity to overcome any vocabulary difficulties while processing their online texts or answering reading questions. The fact that visual information ranked high in their preference strengthens the view that visual information, which is easily presented via computers, could be supportive throughout the reading comprehension process.

| | Strongly disagree | | Disagree | | Not sure | | Agree | | Strongly agree | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Statements** | **Freq.** | **%** | **Freq.** | **%** | **Freq.** | **%** | **Freq.** | **%** | **Freq** | **%** | **SD** |
| **16** | 1 | 2.5 | 1 | 2.5 | 4 | 10 | 28 | 70 | 6 | 15 | .98 |
| **17** | 1 | 2.5 | 1 | 2.5 | 22 | 55 | 12 | 30 | 4 | 10 | .80 |
| **18** | 2 | 5 | 6 | 15 | 8 | 20 | 14 | 35 | 10 | 25 | .82 |
| **19** | 1 | 2.5 | 1 | 2.5 | 3 | 7.5 | 15 | 37.5 | 20 | 50 | .88 |
| **20** | 1 | 2.5 | 2 | 5 | 2 | 5 | 22 | 55 | 13 | 32.5 | .92 |

**Table 1**. Student attitudes towards online reading classes

On the other hand, the lowest frequency score (Strongly Agree: 10%) was for Statement 17: In EFL reading courses, listening to the written text helps me comprehend it better. This can be partly explained by the fact that when a text was difficult for readers to comprehend, audio support did not facilitate comprehension as it was of no support to learners' lexico-grammatical problems.

## 4 Concluding remarks

Without doubt, it has repeatedly been stated that the use of open online technologies can helps foreign language learners enhance their reading, writing, listening and speaking skills while practising their critical thinking and collaboration skills through their participation in online discussion forums (Archambault & Crippen, 2009; Gregor & Cuskelly, 1994; Yang, 2009). In agreement with previous studies, the findings of the present small-scale research further support the view that students seem to have a positive attitude towards integrating CALL into their learning. According to Ayres (2002, p. 247), "learners appreciate and value the learning that they do using the computers"; similarly, Bulut and AbuSeileek (2006) also reported highly positive attitudes towards online English language learning. Moreover, following the claim that "students should like and favor the subject or the activities in the learning environment in order to develop positive attitudes toward learning" (Almahboub, 2000, p. 66), the findings of the present study suggest that as intermediate Greek EFL learners' attitudes were generally positive, they enjoyed the computer-based activities designed.

The current study has, nevertheless, presented a number of challenges and limitations, especially regarding the Student Attitude Questionnaire. While it has provided useful insights into learners' perceptions of online classes, we must be wary of the limitations of young student-opinion data as, at best, these data indicated trends in perceived strategy use. Moreover, other attitudes that were not included in the questionnaire might have been present, or even that the reported ones might have been used more or less often than participants indicated. The fact that a large number of responses were collected, following a standardised set of procedures, could, however, add to the validity of the present findings.

Finally, the online class components and the set of pertinent online activities used in the present study complemented the learning objectives of the actual curriculum, while exposing learners to authentic material and engaging them in real-life tasks.

## Further Reading

Atkins, R. (1991). *Distance Education: New Technologies and Opportunities for Developing Distance Education in New South Wales*. New South Wales: New South Wales Education Department.

Bahman, G., Seyyed, M., Kamal, K., Parviz, A., & Alireza, H. (2011). The impact of asynchronous computer-assisted language learning approaches on English as a foreign language high and low achievers' vocabulary retention and recall. *Computer-Assisted Language Learning, 24*(5), 383–391.

Bax, S. (2003). CALL - past, present and future. *System, 31*(1), 13–28.

Chapelle, C. (2001). *Computer Applications in Second Language Acquisition*. Cambridge: Cambridge University Press.

Compton, L. (2009). Preparing language teachers to teach language online: a look at skills, roles and responsibilities. *Computer-Assisted Language Learning, 22*(1), 73–99.

Golonka, E., Bowles, A., Frank, V., Richardson, D., & Freynik, S. (2014). Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer-Assisted Language Learning, 2*(1), 70–105.

Simsek-Sagin, C. (2008). Students’ attitudes towards integration of ICTS in a reading course: A case in Turkey. *Computers & Education, 51*(1), 200–211.

Stephenson, J. (2001). *Teaching & Learning Online: Pedagogies for New Technologies*. Abingdon, Oxon: Routledge Publications.

Stickler, U., & Hauck, M. (2006). What does it take to teach online? Towards a pedagogy for online language teaching and learning. *CALICO Journal, 23*(3), 463–475.

Tuncok, B. (2010). *A Case Study: Students’ Attitudes towards Computer Assisted Learning, Computer Assisted Language Learning and Foreign Language Learning*. Unpublished Master's thesis, Middle East Technical University.

Walther, J. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research, 19*(1), 52–90.

Walther, J. (1995). Relational aspects of computer-mediated communication: Experimental observations over time. *Organizational Science, 6*(2), 186–203.

Walther, J. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction*. Communication Research, 23*(1), 3–43.

White, C. (2003). *Language Learning in Distance Education*. Cambridge: Cambridge University Press.

Yu-Fen, Y. (2011). Engaging students in an online situated language learning environment. *Computer-assisted Language Learning, 24*(2), 181–198.

## References

Almahboub, S. (2000). *Attitudes toward Computer Use and Gender Differences among Kuwaiti Sixth-Grade Students*. Unpublished PhD thesis, University of North Texas.

Archambault, L., & Crippen, K. (2009). K-12 distance educators at work: Who's teaching online across the United States. *Journal of Research on Technology and Education, 41*(4), 366–387.

Ayres, R. (2002). Learner attitudes toward the use of CALL. *Computer-Assisted Language Learning, 15*(3), 241–249.

Bulut, D. & AbuSeileek, A. F. (2006). Learner attitude toward CALL and level of achievement in basic language kills. *Journal of Institute of Social Sciences of Erciyes University, 23*(2), 112–129.

Felix, U. (2003). An orchestrated vision of language learning online. In U. Felix (Ed.), *Language Learning Online: Towards Best Practice* (pp. 7–20). Lisse: Swets & Zeitlinger.

Furstenberg, G. (1997). Teaching with technology: What is at stake? *ADFL Bulletin, 28*(3), 21–25.

Gillespie, J. & McKee, J. (1999). Does it fit and does it make any difference? Integrating CALL into the curriculum. *Computer-Assisted Language Learning, 12*(5), 441–455.

Greenfield, R. (2003). Collaborative e-mail exchange for teaching secondary ESL: A case study in Hong Kong. *Language Learning & Technology, 7*(1), 46–70.

Gregor, S. & Cuskelly, E. (1994). Computer mediated communication in distance education. Journal of *Computer-Assisted Learning, 10*(3), 168–181.

Kelm, O. (1998). The use of electronic mail in foreign language classes. In J. Swaffar, S. Romano, P. Markley, & K. Arens (Eds.) *Language Learning Online: Theory and Practice in the ESL and L2 Computer Classroom* (pp. 141–153). Austin: Labyrinth Publications.

Lasagabaster, D., & Sierra, J. (2003). Students' evaluation of CALL software programs. *Educational Media International, 40*(3-4), 293–304.

Levinsen, K. (2006). Collaborative online teaching: The inevitable path to deep learning and knowledge sharing? *The Electronic Journal of e-Learning, 4*(1), 41–48.

McIntyre, S., Mirriahi, N., & Watson, K. (2014). *Why is online teaching important? Learning to Teach Online Course Material*. New South Wales: University of New South Wales Massive Open Online Course Material.

Parker, P., Maor, D., & Herrington, J. (2013). Authentic online learning: Aligning learner needs, pedagogy and technology. *Issues in Educational Research, 23*(2), 227–241.

Warschauer, M. & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching, 31*, 57–71.

Wiersma, W. (2008). *Research Methods in Education: An Introduction*. Needham Heights: Allyn & Bacon.

Yang, S. (2009). Using blogs to enhance critical reflection and community of practice. *Educational Technology & Society, 12*(2), 11–21.

# Integrating Technology with Language Assessment: Automated Speaking Assessment

**Graham Seed**, Cambridge English Language Assessment, United Kingdom
**Jing Xu**, Cambridge English Language Assessment, United Kingdom

**Abstract:** This paper evaluates a computer-delivered, fully automated speaking test of English that may be used for selection, placement, or end-of-course evaluation of students at higher education institutions. It begins with a review of the test, including its structure and task types, and then summarises the validity evidence that has been obtained to support the use of an auto-marker. Approximately 2,500 candidates of various first languages and English proficiency levels sat the test and completed a post-test survey. Their oral responses were recorded by an online test delivery system and evaluated by both the auto-marker and certified human examiners. Some validity enquiries were made about the construct coverage of the auto-marker, the agreement and relationship between automated scores and human scores and candidates' perceptions of a non-human examiner. The research reported in this paper sheds light on future development of automated speaking assessment.

## 1 Introduction

With the advancement of natural language processing, machine learning, and speech recognition technologies, automated evaluation of non-native English speech is no longer a dream. An auto-marker, once accurately calibrated, is expected to enhance the reliability of a test because it eliminates idiosyncratic behaviours of human raters (Brown, 2012). In addition, implementing an auto-marker for a speaking test of English would significantly shorten the turnaround time of score reporting and reduce the cost spent on examiner training and hiring.

This is a relatively new and innovative approach to marking Speaking, which has only been researched in a small number of studies so far, for example, Bernstein, Van Moere and Cheng (2010), Chapelle and Chung (2010), and Xi, Higgins, Zechner and Williamson (2008).

Such an auto-marker has begun to be devised for the speaking component of a new product from Cambridge English Language Assessment called Linguaskill. This short paper begins with a review of the speaking test, including its structure and task types, and then summarises the validity evidence that has been obtained so far to support potential use of the auto-marker.

## 2 Market requirements for Linguaskill

In preparation for creating the Linguaskill suite of tests, market research was carried out to perform a needs analysis for a test of English that is used for selection, placement, or end-of-course evaluation of students enrolled at post-secondary educational institutions, or alternatively for recruitment or progress in employment. Key findings were that it should be a test of general English ability, able to test a range of language levels, and that the test should be flexible enough to be delivered on-demand, with no need to register candidates in advance of the test sitting, and with accurate results delivered as soon as possible.

## 3 The Linguaskill speaking test

The Linguaskill speaking test can be taken on any computer with a fast internet connection. No software needs to be downloaded or installed. The test is approximately 15 minutes long and contains 5 parts, 4 of which elicit free speech responses and 1 (Part 2) contains sentences shown on screen to be read aloud. The 5 parts progress from the very accessible to the more challenging, and aim to cover as much of the construct of speaking as possible within the constraints of a single-candidate online format. Of the free speech parts, Part 1 requires the candidate to answer eight simple questions about their everyday life; Part 3 is a presentation about a given topic; Part 4's purpose is to leave an answerphone message with information based on authentic graphical input, and Part 5 is to give opinions on five questions related to a given scenario. For the latter three parts, candidates are given an amount of preparation time before the response time.

Linguaskill reports test outcomes based on the Common European Framework of Reference for Languages (CEFR) scale from A1 to C1 or above. In addition, Linguaskill reports an extra proficiency level, called Below A1 which indicates minimal language ability. A candidate's final score is the average of the scores received in each part of the test. This composite score will eventually be converted into a CEFR level according to predetermined cut-off scores for each proficiency level.

## 4 The auto-marker

In terms of fulfilling the market requirements described in section 2, Linguaskill's speaking component is manifested in an online computer-delivered and auto-marked test, powered by cutting-edge natural language processing (NLP) and machine learning technologies. That is, all the oral responses that candidates produce in the tests are not graded by human examiners but by a speech auto-marker. The auto-marker is essentially a series of computer algorithms that learned how to mark test responses from a large collection of learner responses that had already been marked by human experts.

Speech is captured during the response periods within the test and sent to the auto-marker, which can extract non-content features such as fluency, hesitation, stress patterns and other pronunciation features. Concurrently, speech recognition software transcribes input, from which certain features of grammar and vocabulary accuracy and complexity, along with elements of content relevance based on the input prompt, are extracted. An algorithm then evaluates these features and also converts the evaluations to a single score which is sent back to the test platform to create score reports for the customer.

## 5 Research

A large-scale trial of the Linguaskill speaking test was carried out from December 2016 to February 2017 with the following research questions (RQs) in mind:

(1) RQ1: How well did the auto-marker agree with human raters in scoring test responses?
(2) RQ2: Do the five test parts measure a similar speaking construct or different ones?
(3) RQ3: What were learners' perceptions of the automated speaking test?

### 5.1 Methodology

Test-takers took one of two pre-validated versions of the Linguaskill speaking test, delivered online, which recorded their responses to the tasks within the five parts of the test. These responses were sent both to the auto-marker and to certified human raters, who returned a global mark per part based on a six point scale. The part responses were automatically and randomly allocated to two or more human raters so that one candidate’s responses were never marked by only one sole human rater. This increased the reliability of the human marking as, due to the high cost of human marking and the large number of participants in this trial, it would have been unfeasible to double-mark all candidate responses.

Test-takers were also asked to complete a post-test survey about their experiences during the test.

### 5.2 Participants

Data was collected from 2,612 English-language learners from 23 different countries who participated in the trial, taking both the Linguaskill speaking test and completing the post-test survey. A majority of them were recruited from Brazil (27.2%), India (25.2%) and Japan (13.7%), with the largest language groups being Portuguese (27.1%), Hindi (23.7%), Japanese (13.4%) and Spanish (11.5%). Test-takers aged between 16 and 24 accounted for 71.8% of the sample, and were split roughly half-half between male and female. Measured using the human raters’ overall scores, the distribution of test-takers’ overall CEFR scores showed a classic bell curve, with the majority being at B1 or B2 level (67.99%).

### 5.3 Findings

#### *5.3.1 Human-machine agreement*

In this study, human expert judgment was used as a gold standard for assessing the accuracy of automated scores. For this reason, the reliability of human marking had to be first examined to ensure that the gold standard was not fallible. The reliability of human marking was estimated using intraclass correlation coefficients (ICCs) that indicate the correlation between a single rating on a response and other ratings on the same response (Shrout & Fleiss, 1979, p. 422). To perform this analysis, five raters were randomly chosen from the 31 raters who had marked the trial responses; they were then asked to mark the same ‘common set’, a subset of data that contained responses produced by 60 of the candidates. These candidates were randomly drawn from the large pool of Linguaskill trial participants and represented English-language learners of various levels of oral proficiency. Because both the raters and candidates were randomly selected from a larger population and each candidate was marked by the same five raters, two-way random ICCs were computed using SPSS (Shrout & Fleiss, 1979, p. 421).

The magnitude of the ICCs of human ratings on each part varies from 0.84 to 0.91, and the coefficient of the average of the part scores (which is used as the final test score) is 0.91. This indicates excellent reliability of single human rating on each part and on the whole test, as an ICC above 0.75 suggests excellent rater reliability (Cicchetti, 1994). This finding served as the premise for using single human marking as a gold criterion for evaluating the auto-markers' performance.

Spearman's rank order correlation coefficients were then calculated between all the marks awarded by human raters and those of the auto-marker. The findings by part were: Part 1 – 0.61 ($p < .01$), Part 2 – 0.54 ($p < .01$), Part 3 – 0.66 ($p < .01$), Part 4 – 0.61 ($p < .01$) and Part 5 – 0.69 ($p < .01$). The overall correlation was 0.80 ($p < .01$), suggesting that all the parts of the test are needed to present an accurate indication of a candidate's speaking ability.

At the same time, the level of agreement between the overall CEFR level awarded by the human raters for one candidate and that awarded by the auto-marker, was calculated. This showed that the human raters and auto-marker agreed exactly in 41.4% of cases; furthermore, agreement at either the same or one adjacent CEFR level happened in 89.3% of cases. Thus only 10.7% of candidate results were misclassified where the difference was two or more CEFR levels. When a sample of the candidates that fell into this category were listened to, it was noted how a large proportion of these recordings suffered from audio quality issues, such as a 'fuzzy' background noise, suggesting that the auto-marker may currently not be so able to give an accurate result in these conditions, and therefore more work should be done on reducing the circumstances which create problems with the audio quality of the recordings.

### *5.3.2 Factorial structure of the test*

The factor loadings of each of the five parts to a single latent speaking construct (a single-trait model) were estimated using confirmatory factor analysis to check the factor structure of the test. Findings showed that a single-trait model in which the five parts are loaded onto a general speaking factor best fits the data; in addition, the factor loadings were not markedly different except the one for Part 2 (the reading aloud task). This suggests that the test as a whole assesses the general speaking ability whereas the speaking skill assessed by Part 2 is slightly different from that assessed by other spontaneous speaking tasks.

### *5.3.3 Participants' perceptions*

Participants' perceptions of the test content, the test environment and general attitudes to auto-marking were gathered using a post-test online survey. On the whole, participants felt positive (46.6%) or very positive (13.8%) towards the speaking test.

A number of other likert-scale questions were asked. Positive perceptions were registered about the test content. For example, the percentage of respondents who agreed or strongly agreed that the test instructions were clear was 85.6%; 79.2% agreed or strongly agreed that the visuals in the test were clear and understandable; and 69.5% agreed or strongly agreed that the test allowed them to properly demonstrate their English speaking ability. A representative comment from open-ended responses about the test content was:

> I find the topics relevant, not too easy nor difficult. I think that these topics are related to what normally happens in daily life. These are topics that most people learning English should master because they are what takes place in the real world.

About the test format, one respondent said:

> I always feel worried in exams, but as I hear the questions I felt more comfortable and relax [sic]. The speaking test was developed gradually, so you feel good when you notice that you start with a repetition, and then you answer easy questions, and then you have to think a little more to answer the final questions.

Other notable findings were that 34.0% of respondents indicated they they suffered from other noise in the test room, and 39.2% encountered some sort of technical issues. In addition, 41.9% claimed to be nervous or worried before or during the test.

Some test-takers found the experience of speaking to a computer positive:

> I like the new experience to talk with a computer, I felt less pressure than talking with a person.
>
> I felt free to talk to a computer just as if I was talking to a real person.

However, others were less confident:

> I got little nervous because I cannot see a face. This system is efficient but little lonely.
>
> It could be a little bit easier and human interaction is lost and it is important when you use English in real life.

This was often the case when thinking about the test being auto-marked:

> Somehow, I felt impelled to speak in a very mechanical way, as I was worried if the corrector would understand the recording properly.

On the other hand, 64.8% of respondents said that they would not worry about the fact the test would be auto-marked.

## 6 Conclusions and implications

The trial showed that, on the whole, there is evidence to support a claim that the auto-marker's scoring accuracy is satisfactory, and that it is more reliable than a human rater, in that the former does not suffer from negative affective factors such as tiredness. Nevertheless, to be even more accurate, the auto-marker could continue to be trained using more data from candidates from different L1s and at different ability levels. It was also seen how poor quality of audio recording affected the accuracy of the auto-marker. Therefore, work can be done on making the test experience better for the candidates to ensure appropriate recording.

Using an auto-marker for online speaking tests is an innovative solution to the market need, and use of this type of technology in this field is still relatively in its infancy. Because of test-takers' lack of experience with such methods, survey responses have also shown the need for a greater test-taker acceptance of taking auto-marked speaking tests online.

Nevertheless, the future potential of auto-marked online speaking tests is judged to be vast, with the accelerated development of this technology a major factor to continue improving the accuracy of results and the experience of taking such a test.

## References

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests*. Language Testing, 27*(3), 355–377.

Brown, A. (2012) Interlocutor and rater training. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 413–425). Oxford: Routledge.

Chapelle, C. A. & Chung, Y-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing, 27*(3), 301–315.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. *Psychological Assessment, 6*(4), 284–290.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated Scoring of Spontaneous Speech Using SpeechRater(SM) v1.0*. Princeton: Educational Testing Service.

# Using Technologies for Teaching a Subject with CLIL: A Look at Teachers' Reactions

**Letizia Cinganotto**, INDIRE, Italy
**Juliet Wilson**, Cambridge English, UK

**Abstract:** This paper begins with a brief outline of the latest research in the field of technology-enhanced language learning and CLIL (Content and Language Integrated Learning). It goes on to describe the main outcomes from a questionnaire delivered by the authors to 241 teachers from across the world attending the free MOOC "Teaching your subject in English" by FutureLearn and Cambridge English. The questionnaire was aimed at investigating the teachers' reactions to CLIL programmes and then more specifically to the use of technologies for teaching a subject in English in order to ascertain what added value web and digital media contributed to different aspects of the teaching and learning process.

## 1 Introduction

21st century learners are constantly exposed to technology in their daily life: smartphones, mobile devices and social networks are now the most common and natural way our students communicate with their friends and with the world.

Gardner and Davis (2014) define digital learners as "the App Generation": the current generation who are so deeply immersed in digital media, with all their inherent strengths and weaknesses. These media can foreclose a sense of identity, encourage superficial relations with others and stunt creative imagination; but on the other hand, they can also promote a strong sense of identity, allow deep relationships, and stimulate creativity. In particular, according to Gardner and Davies, the impact of multimedia can be visible in three vital areas of adolescent life: identity, intimacy and imagination. The challenge, especially for educators, is to use the potential of apps and digital media as a springboard to greater creativity and higher aspirations.

We may need to rethink our teaching practices and strategies to meet students' learning needs and individual styles and to exploit the communication channels which are more familiar to them. In other words, we may need to interweave formal, informal and non-formal learning pathways.

So, what are the benefits of teaching a subject in English using learning technologies? What are teachers' reactions and perceptions of technology-enhanced language learning and CLIL?

These were the main questions underlining the research project carried out by the authors and described in this paper. We wanted to investigate how teachers really feel and react to the use of technologies for CLIL in English.

## 2 Teaching CLIL with technologies: background

In foreign language teaching it is argued that it is becoming increasingly important to develop not only students' "communicative competence" (Canale & Swain, 1980), but also their "electronic communicative competence" (Simpson, 2005) or "ICT competence" (Walker, 2007).

Technologies can have a huge impact on language learning and on CLIL (Coyle, Hood, & Marsh, 2010; Vlachos, 2009), for example, students' active participation and self-expression; opportunities for authentic language use; students' collaboration and socialization and working across the curriculum (Singhal, 1997, Warschauer & Whittaker, 1997).

Recent developments in language learning and teaching with technologies have focused on some new research areas: CALL (Computer Assisted Language Learning) (Chapelle, 2001), which sees technology as assisting language learning; TELL (Technology Enhanced Language Learning) (Bush and Terry, 1997; Walker & White, 2013), which refers to ICT as a part of the environment in which language exists and is used and provides not only new tools, but also new educational contexts and settings; WELL (Web-enhanced Language Learning), referring to the internet as a medium for instruction; NBLL (Network-based Language Learning), (Warschauer & Kern, 2000), underlining the interconnectivity of computers in facilitating interpersonal digital communication; and MALL (Mobile Assisted Language Learning) (Kukulska-Hulme & Shields, 2008; Vavoula, Pachler, & Kukulska-Hulme, 2009), focusing on the use of personal, portable devices to enable new ways of learning, across different contexts of use.

In 2014 The European Commission published a report entitled "Improving the effectiveness of language learning: CLIL and computer assisted language learning", which highlights the link between language learning, CLIL and technologies. In particular the following options are mentioned:

- authentic foreign language material, such as video clips, flash-animations, web-quests, podcasts
- online environments, social media, or voice/video conferencing
- language-learning tools (online apps or software)
- online proprietary virtual learning environments
- game-based learning.

The CLIL approach is gaining ground across Europe, as shown in the Eurydice report "Keydata on Teaching Languages at school in Europe" (2017). In Italy it was introduced as mandatory at secondary level by Decrees 88/89 dated 2010: The "Good School Reform" (Law 107/2015) encouraged the introduction of CLIL from primary school up, highlighting the positive impact this methodology can have on the internationalization of school curricula and on students' learning outcomes.

Within this framework, Continuous Professional Development for teachers, as far as both language learning/CLIL and digital dimensions are concerned (Cinganotto & Cuccurullo, 2016), has become crucial.

The National Teacher Training Plan launched by the Italian Ministry of Education in October 2016 mentions language learning, CLIL and multimedia competences among the top priorities for teacher training in the next few years.

Against this background, we wanted to try to better understand teachers' perceptions on the use of technologies for teaching a subject in English with CLIL.

## 3 The context

A questionnaire was delivered to a sample of teachers attending a free MOOC entitled "Teaching your subject in English" promoted by FutureLearn and Cambridge English. The questionnaire was planned by the authors of this paper and the outcomes were presented during ALTE conference 2017 in Bologna.

The MOOC "Teaching your subject in English" aims to build teachers' confidence in using English effectively to teach subjects such as Maths, Science and History. The 5-week language course provides a range of functional language and practical tips for teaching a subject in English. Teachers can join a community of teachers from across the world, share their expertise and look for new ideas from a range of different contexts. There are also opportunities to explore digital tools for teaching and learning. The course is designed for secondary school subject teachers who deliver lessons in English, involved in Bilingual Education or CLIL.

## 4 The sample of respondents

241 teachers answered the questionnaire: the largest nationality represented was Italian (26.8%) followed by Spanish (18%); the other respondents were from a number of other countries across the world, for example UK, Russia, Mexico, Portugal, Kazakhstan, Colombia. They were mainly teaching in state schools (53.5%). 35.5% of them worked in schools which provided an international curriculum. 64% of the participants were upper secondary school teachers, with 11 or more years of working experience. The majority of respondents taught English as a foreign language (73.3%), but there were also some teachers of Italian (11.8%). Their level of English according to the Common European Framework of Reference for Languages (CEFR) was generally high: C2 (36.4%), C1 (32.5%), and B2 (26.3%).The other lower levels were distributed over the remaining small percentages.

39.5% of the respondents had more than 10 years' experience in teaching a subject in English and more than 50% of them had attended specific training on CLIL in English, in particular through online training initiatives such as MOOCs, webinars and online courses. A very similar percentage (53.3%) of them had also attended training initiatives on using technologies for language learning and CLIL in English via online or blended training courses.

The majority of the respondents stated they had excellent (14.9%), very good (31.6%) or good (33.8%) competences in using technology or teaching – a very encouraging figure, showing positive attitudes towards learning technologies for CLIL.

## 5 Main findings from the questionnaire

We were interested in finding out teachers' perceptions of the value of CLIL programmes in terms of different learning outcomes. The possible response range was from 1 (negative) to 5 (positive). "Language learning outcomes'' scored highest (m = 4.3), "Other learning outcomes'' such as transversal or soft skills were also rated positively (m = 4.1) with “Subject learning outcomes’’ scoring 4.0. These data show the general positive perceptions about the potential of CLIL as an added value to the curriculum, as some of the comments from the teachers collected though the questionnaire confirm:

> My students are more motivated to learn English and the other subject.
>
> The CLIL approach helps to present the subject matter in a simple way.
>
> Students are more motivated, focusing on context rather than language.
>
> Invisible language learning!
>
> It gives exposure to the language of instruction and to the use of the language of instruction.

Teaching a subject in English is considered an innovative approach which allows students to have full immersion in the foreign language through the learning of the content and full immersion in the content through the learning of the foreign language: mutual beneficial effects with strong motivation and engagement for the students.

Our other main interest was to understand teachers’ use and perception of technologies. We asked how often teachers used digital tools and resources in their teaching. The answers were very encouraging: 42.1% answered “often” and 31.1% answered “always”. These data show that technologies for teaching CLIL are now well-embedded in teachers’ repertoires and it is clear that teachers are generally aware of the importance of the use of digital tools and resources for enhancing students’ motivation and learning outcomes.

To the question “Do you use classroom material from the internet?” 93% provided a positive answer: a very high figure which confirms the assumption that the internet has now become an integral part of lesson planning.

The use of technology was one area of investigation but we were also interested in perceived value: “When you teach a subject in English, how much value do you think technology adds to the following aspects?” (1 = technology adds no value, 5 = technology adds a lot of value)

The highest value (m = 4.5) was registered for listening activities. Audio oral skills are often complicated to practise in the classroom as these skills need specific support. Technology can help for example, with the use of videos, podcasts, and movie clips etc. that can be used within an ESL or a CLIL activity. This combines interest and fun with the “invisible language learning” mentioned by one of the respondents. Examples specifically mentioned were: YouTube videos and tools; BBC podcasts (bbc.co.uk/podcasts); Talkgroups (eg. voxopop.com) and recording tools (eg. vocaroo.com).

The second highest value was for students’ presentation (m = 4.3): a wide range of teaching strategies and techniques can take advantage of technology to share students’ individual or group work with the teachers and with their peers, such as webquests, TBLL (Task-based Language Learning), and PBL (Project-based Learning). Technologies can help create multimedia products as individual or collective and cooperative work to be shared in class or to be assessed by the teachers or by their peers.

Vocabulary activities also scored highly (m = 4.2): there is a wide range of online dictionaries, translating tools and other web tools which are specifically designed to practise and extend vocabulary. An example of these tools, also mentioned by some respondents, is Snappy Words (snappywords.com), which can generate high impact semantic and grammatical mind maps starting from a certain key word given.

Many other options also scored well: collaboration and communication with and between learners (m = 4), lesson planning (m = 3.9), reading activities (m = 3.9), assessment (m = 3, 8).

The lowest score was for writing activities (m = 3.6). This could be that in order to develop writing skills in English, teachers prefer "traditional" pen and paper. However, there is a wide range of tools designed to help students practise their writing skills and some of these were mentioned by the respondents: Storybird (storybird.com) for digital story-telling; Write&Improve by Cambridge English (writeandimprove.com) for improving writing skills; and tools for blogs or wikis (eg. Wordpress, Wikispace).

The picture below shows the added value perceived by the respondents with reference to the different aspects of the teaching and learning process, provided as options (Figure 1):

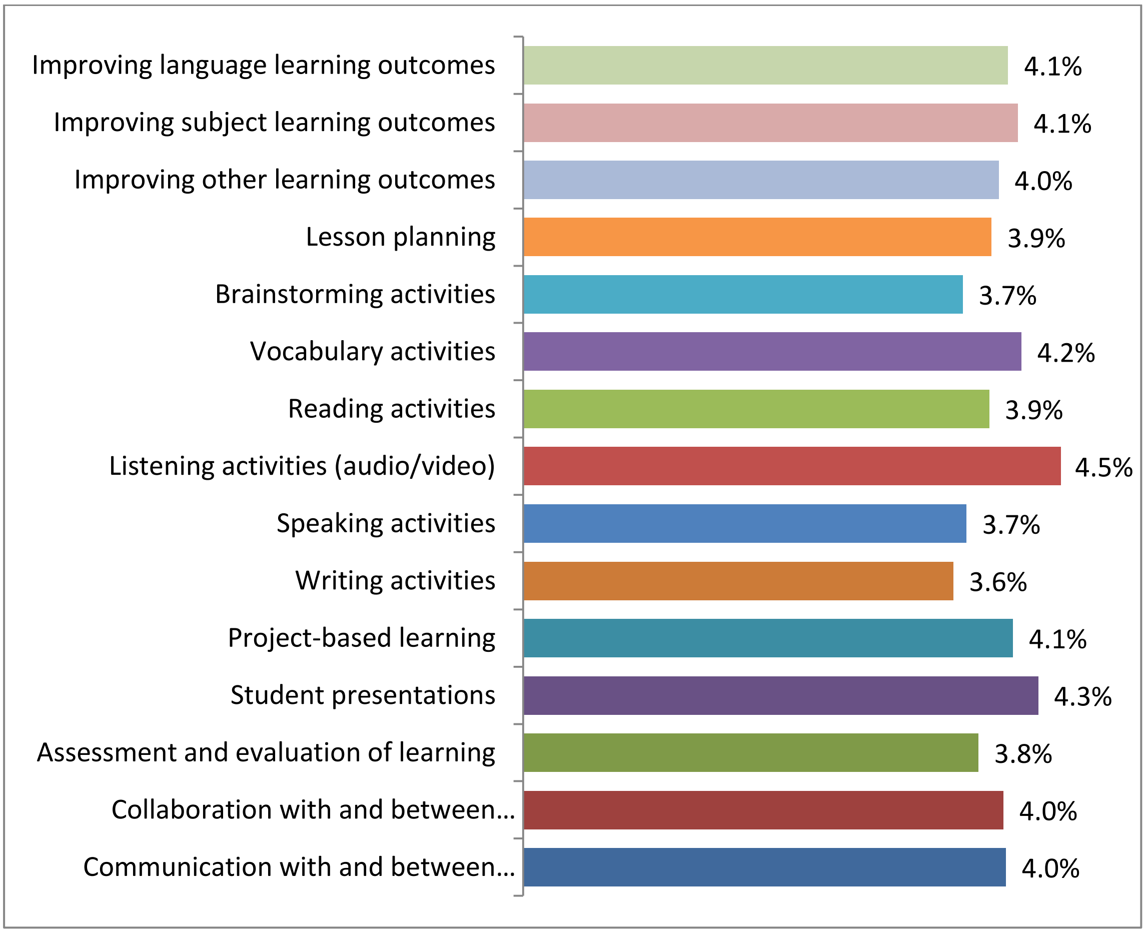


**Figure 1.** Respondents' views on aspects of teaching

Some of the comments from respondents highlight key areas:

> Technology allows more focused teaching tailored to students' strengths and weaknesses with the flexibility of 'anytime, anywhere' access, opportunities to collaborate on assignments with people outside or inside school.
>
> Digital tools and resources are an important addition and improvement to teaching and learning. But they have to be chosen and used correctly and constantly. Being a digital teacher is not easy and cannot be an off-the cuff competence.
>
> Always remember that they are only tools; their benefits depend on the context the teacher operates in.

## 6 Conclusions

The research project was designed to investigate the added value of technologies in teaching a subject in English according to the perceptions of a sample of teachers (241) attending a MOOC by FutureLearn and Cambridge English.

After a brief literature review on the main research trends relevant to language learning and CLIL with the use of technologies, a questionnaire was delivered to the teachers who had enrolled in the first version of the free MOOC "Teaching your subject in English".

The teachers who responded had a high level of awareness of using technologies in teaching a subject in English as they have already attended other online training initiatives both on CLIL and on learning technologies.

Their perceptions and feelings towards the use of technologies were generally positive: they regularly access the internet for materials and ideas to use during their lessons; they think that technologies can have a strong impact on their teaching practices and on the development of the students' skills, in particular audio oral skills, vocabulary skills and students' presentations.

As the literature shows, the use of multimedia and digital tools for language learning and CLIL has huge potential. However this potential must be realised by the teacher who continues to have an indispensable role as a facilitator of learning processes and knowledge building.

## Acknowledgements

The authors wish to thank INDIRE and Cambridge English for supporting the research design.

## References

Bush M. D. & Terry R. M. (Eds.) (1997). *Technology-enhanced language learning*. Lincolnwood: National Textbook Company.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Chapelle C. (2001). *Computer Applications in Second Language Acquisition*. Cambridge: Cambridge University Press.

Cinganotto, L. & Cuccurullo, D. (2016). CLIL and CALL for a teacher's expertise: An international training experience. *Form@re - Open Journal per la formazione in rete, [S.l.], 16*(2), 319–336.

Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and Language Integrated Learning*. Cambridge: Cambridge University Press.

Eurydice (2017). *Keydata on Teaching Languages at school in Europe*. Retrieved from: https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/images/0/06/KDL_2017_internet.pdf.

FutureLearn and Cambridge English MOOC (no date) *Teaching your subject in English.* Retrieved from https://www.coursetalk.com/providers/futurelearn/courses/teaching-your-subject-in-english

Gardner H. & Davis K. (2014). *The App Generation*. London: Yale University Press.

Kukulska-Hulme A. & Shield L. (2008). An overview of mobile-assisted language learning: From content delivery to supported collaboration and interaction. *ReCALL, 20*(3), 249–252.

Simpson, J. (2005). Learning electronic literacy skills in an online language learning community. *Computer Assisted Language Learning, 18*(4), 327–345.

Singhal, M. (1997). The Internet and foreign language education: Benefits and challenges. *The Internet TESL Journal*, *3*(6). Retrieved from http://iteslj.org/Articles/Singhal-Internet.html

Vavoula G., Pachler N., & Kukulska-Hulme A. (Eds.) (2009). *Researching mobile learning: Frameworks, tools and research designs*. Oxford: Peter Lang Verlag.

Vlachos, K. (2009). The potential of information communication technologies (ICT) in content and language integrated learning: The case of English as a second/foreign language. In M. Ruiz-Garrido, I. Fortanet-Gómez, D. Marsh, P. Mehisto, D. Wolff, R. Aliaga, T. Asikainen, M.J. Frigols-Martin, S. Hughes, G. Lange, M. Ruiz, I. Morales-Gómez, G. R. Aliaga, & M. Frigols (Eds.), *CLIL practice: perspectives from the field* (pp. 189–198). Juvaskyla: University of Jyvaskyla.

Walker, A. & White, G. (2013). *Technology Enhanced Language Learning. Connecting Theory and Practice*. Oxford: Oxford University Press.

Warschauer, M. & Whittaker, P. F. (1997). The Internet for English teaching: Guidelines for teachers. *The Internet TESL Journal, 3*(10), 27–33.

Warschauer M. & Kern R. (2000). *Network-based Language Teaching: Concepts and Practice*. Cambridge, Cambridge University Press.

European
Commission

Under the auspices of the Secretary General of the Council of Europe, Mr Thorbjørn Jagland

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE